Monographs
on Statistics and
Applied Probability 79

Bayesian Methods for Finite Population Sampling

Malay Ghosh and Glen Meeden

Springer-Science+Business Media, B.V.

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

D.R. Cox, V. Isham, N. Keiding, N. Reid and H. Tong

1 Stochastic Population Models in Ecology and Epidemiology *M.S. Bartlett* (1960)
2 Queues *D.R. Cox and W.L. Smith* (1961)
3 Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
4 The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
5 Population Genetics *W.J. Ewens* (1969)
6 Probability, Statistics and Time *M.S. Bartlett* (1975)
7 Statistical Inference *S.D. Silvey* (1975)
8 The Analysis of Contingency Tables *B.S. Everitt* (1977)
9 Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
10 Stochastic Abundance Models *S. Engen* (1978)
11 Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
12 Point Processes *D.R. Cox and V. Isham* (1980)
13 Identification of Outliers *D.M. Hawkins* (1980)
14 Optimal Design *S.D. Silvey* (1980)
15 Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
16 Classification *A.D. Gordon* (1981)
17 Distribution-free Statistical Methods, 2nd edition *J.S. Maritz* (1995)
18 Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19 Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20 Risk Theory, 3rd edition *R.E. Beard, T. Pentikainen and E. Pesonen* (1984)
21 Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)
22 An Introduction to Latent Variable Models *B.S. Everitt* (1984)
23 Bandit Problems *D.A. Berry and B. Fristedt* (1985)
24 Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
25 The Statistical Analysis of Compositional Data *J. Aitchison* (1986)
26 Density Estimation for Statistics and Data Analysis *B.W. Silverman* (1986)
27 Regression Analysis with Applications *G.B. Wetherill* (1986)
28 Sequential Methods in Statistics, 3rd edition *G.B. Wetherill and K.D. Glazebrook* (1986)
29 Tensor Methods in Statistics *P. McCullagh* (1987)

30 Transformation and Weighting in Regression *R.J. Carroll and D. Ruppert* (1988)
31 Asymptotic Techniques for Use in Statistics *O.E. Barndorff-Nielsen and D.R. Cox* (1989)
32 Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
33 Analysis of Infectious Disease Data *N.G. Becker* (1989)
34 Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35 Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
36 Symmetric Multivariate and Related Distributions *K.-T. Fang, S. Kotz and K.W. Ng* (1990)
37 Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38 Cyclic and Computer Generated Designs, 2nd edition *J.A. John and E.R. Williams* (1995)
39 Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40 Subset Selection in Regression *A.J. Miller* (1990)
41 Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
42 Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
43 Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)
44 Inspection Errors for Attributes in Quality Control *N.L. Johnson, S. Kotz and X. Wu* (1991)
45 The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)
46 The Analysis of Quantal Response Data *B.J.T. Morgan* (1993)
47 Longitudinal Data with Serial Correlation: A State-space Approach *R.H. Jones* (1993)
48 Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)
49 Markov Models and Optimization *M.H.A. Davis* (1993)
50 Networks and Chaos – Statistical and Probabilistic Aspects *O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall* (1993)
51 Number-theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)
52 Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)
53 Practical Risk Theory for Actuaries *C.D. Daykin, T. Pentikäinen and M. Pesonen* (1994)
54 Biplots *J.C. Gower and D.J. Hand* (1996)
55 Predictive Inference: An Introduction *S. Geisser* (1993)
56 Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)
57 An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)
58 Nonparametric Regression and Generalized Linear Models *P.J. Green and B.W. Silverman* (1994)

59 Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)
60 Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)
61 Statistics for Long Memory Processes *J. Beran* (1995)
62 Nonlinear Models for Repeated Measurement Data *M. Davidian and D.M. Giltinan* (1995)
63 Measurement Error in Nonlinear Models *R.J. Carroll, D. Ruppert and L.A. Stefanski* (1995)
64 Analyzing and Modeling Rank Data *J.I. Marden* (1995)
65 Time Series Models – In econometrics, finance and other fields *D.R Cox, D.V. Hinkley and O.E. Barndorff-Nielsen* (1996)
66 Local Polynomial Modelling and Its Applications *J. Fan and I. Gijbels* (1996)
67 Multivariate Dependencies – Models, analysis and interpretation *D.R. Cox and N. Wermuth* (1996)
68 Statistical Inference – Based on the likelihood *A. Azzalini* (1996)
69 Bayes and Empirical Bayes Methods for Data Analysis *B. Carlin and T. Louis* (1996)
70 Hidden Markov and Other Models for Discrete-valued Time Series *I.L. MacDonald and W. Zucchini* (1997)
71 Statistical Evidence - A likelihood paradigm *R. Royall (1997)*
72 Analysis of Incomplete Multivariate Data *J.L. Schafer* (1997)
73 Multivariate Models and Dependence Concepts *H. Joe* (1997)
74 Theory of Sample Surveys *M.E. Thompson* (1997)
75 Retrial Queues *G. Falin and J.G.C Templeton* (1997)
76 Theory of Dispersion Models *B. Jørgensen* (1997)
77 Mixed Poisson Processes *J. Grandell* (1997)
78 Variance Components Estimation - Mixed models, methodologies and applications *P.S.R.S. Rao* (1997)
79 Bayesian Methods in Finite Population Sampling *G. Meeden and M. Ghosh (1997)*

(Full details concerning this series are available from the Publishers).

To

Dola, Debashis and Debadyuti

Nancy, Lisa and Marc

Bayesian Methods for Finite Population Sampling

M. Ghosh

Professor of Statistics
University of Florida
USA

and

G. Meeden

Professor of Statistics
University of Minnesota
USA

Springer-Science+Business Media, B.V.

ISBN 978-0-412-98771-7 ISBN 978-1-4899-3416-1 (eBook)
DOI 10.1007/978-1-4899-3416-1

First edition 1997

Originally published by Chapman & Hall in 1997.
Softcover reprint of the hardcover 1st edition 1997

A Catalogue record for this book is available from the British Library

Printed on permanent acid-free text paper, manufactured in accordance with ANSI/NISO Z39.48 - 1992 and ANSI/NISO Z39.48 - 1984 (Permanence of Paper).

Contents

Preface

The present monograph is primarily an outgrowth of our own research on certain aspects of Bayesian inference in finite population sampling. Finite population sampling has been an integral part of statistics since its beginning. The topic continues its impact in the theory and practice of statistics, especially for researchers in survey sampling.

Inference for finite population sampling utilizes prior information either explicitly or implicitly. Bayesian inference makes explicit use of this information as part of the model. This is in striking contrast to design- based inference in survey sampling where prior knowledge is incorporated only as auxiliary information. On the other hand there is a close relationship between the Bayesian approach and the superpopulation approach, although they differ in their foundational interpretations. Operationally, however, the difference is much less pronounced as many estimators obtained via superpopulation models are also obtainable as Bayes estimators, and vice versa.

This monograph, does not aim to provide a complete up-to-date account of the Bayesian literature in finite population sampling. Rather, it treats the topics reflecting the authors' personal interests. Its main aim is to demonstrate that a variety of levels of prior information can be used in survey sampling in a Bayesian manner. Situations considered range from a noninformative Bayesian justification of standard frequentist methods when the only prior information available is the belief in the exchangeablility of the units to a full-fledged Bayesian model.

This book is primarily for researchers and advanced graduate students working on finite population sampling, but should prove useful also for persons working in Federal, State and other government or non-government agencies. It can also be used as a special topics course in finite population sampling spanning a quarter or a semester.

The current project was initiated in 1992. The authors thank John Kimmel for his advice and encouragement in the initial stage of this project. Since then, we have had the fortune of working with several Chapman Hall Book Editors, and thank them all for their co-operation. Special thanks are due to Mark Pollard at the final stages of the project, and to John Bell, the copy-editor, for his thorough and painstaking job in bringing the manuscript into its final shape.

We gratefully acknowledge the facilities provided by the Department of Statistics, University of Florida, the Departments of Statistics and Biostatistics, University of Minnesota, and the Department of Mathematics and Statistics, Bowling Green State University, where much of the writing was done. Also, the book would not have seen the light of the day without the active support of all our family members.

We thank Bret Presnell who helped a couple of LaTeXnovices get started on the long journey of writing this book. And finally, we thank Charles Geyer for his unfailing and selfless TeXnical advice in the LaTeXpreparation of this manuscript. Without his help we would still be going.

Malay Ghosh — Gainesville, Florida
Glen Meeden — Minneapolis, Minnesota
May 1997

CHAPTER 1

Bayesian foundations

In this chapter we present the underlying foundations of finite population sampling and describe the Bayesian and predictive approach to inferential problems in this area. In Section 1.1 we introduce the notation we will be using throughout the book. In Section 1.2 we restate the work of Basu and Ghosh (1967) which identifies the minimal sufficient statistic for a finite population sampling problem. In Section 1.3 we summarize Basu's work (1969) which describes how the likelihood principle should be applied in finite population sampling. In Section 1.4 we outline the usual Bayesian approach to finite population sampling with particular emphasis on the work of Ericson (1969b). In Section 1.5 we review an approach to finite population sampling which only assumes that the posterior expectation is linear. Finally in Section 1.6 we give a brief outline of the rest of the book.

1.1 Notation

In this section we will introduce the notation which we shall use throughout the book. For now, we will concentrate on the simplest situation in finite population sampling. Let $\mathcal{U}$ denote a finite population which consists of N units labelled $1, 2, \ldots, N$. We will assume that these labels are known and that they often can contain some information about the units. Attached to unit i let y_i be the unknown value of some characteristic of interest. Typically, y_i will be a real number. For this problem $\mathbf{y} = (y_1, \ldots, y_N)^T$ is the unknown state of nature or parameter. $\mathbf{y}$ is assumed to belong to $\mathcal{Y}$, a subset of N-dimensional Euclidean space, $\mathcal{R}^N$. The statistician usually has some prior information about $\mathbf{y}$ and this could influence the choice of $\mathcal{Y}$. However, in most cases convenience and tradition seem to dictate the choice of the parameter space and usually $\mathcal{Y}$ is taken to be $\mathcal{R}^N$. In what follows we will sometimes assume that the parameter space is a finite subset of $\mathcal{R}^N$ of a par-

ticular special form. If $\mathbf{b} = (b_1, \ldots, b_k)^T$ is a k-dimensional vector of distinct real numbers then we let

$$\begin{aligned} \mathcal{Y}(\mathbf{b}) &= \{\mathbf{y} : \text{such that for } i = 1, \ldots, N,\ y_i = b_j \\ &\qquad \text{for some } j = 1, \ldots, k\}. \end{aligned} \tag{1.1}$$

In many problems in finite population sampling there are additional characteristics or variables associated with each unit which are known to the statistician. For unit i let x_i denote a possible vector of other characteristics, all of which are assumed to be known. We let $\mathbf{x}$ denote the collection of these vectors for the entire population. Hence in the usual frequentist theory the x_i's and their possible relationship to the y_i's summarize the statistician's prior information about $\mathbf{y}$.

A subset s of $\{1, 2, \ldots, N\}$ is called a sample. Let $n(s)$ denote the number of elements belonging to s. Let S denote the set of all possible samples. A (nonsequential) sampling design is a function p defined on S such that $p(s) \in [0, 1]$ for every nonempty $s \in S$ and $\sum_{s \in S} p(s) = 1$. Given $\mathbf{y} \in \mathcal{Y}$ and $s = \{i_1, \ldots, i_{n(s)}\}$, where $1 \leq i_1 < \cdots < i_{n(s)} \leq N$ let $\mathbf{y}(s) = (y_{i_1}, \ldots, y_{i_{n(s)}})^T$.

Given a parameter space $\mathcal{Y}$ and design p a typical sample point is the set of labels of the units contained in the observed sample along with their values of the characteristic of interest. We will denote such a data point by

$$\begin{aligned} z &= (s, z_s) \\ &= (s, (z_{i_1}, \ldots, z_{i_{n(s)}})^T) \end{aligned} \tag{1.2}$$

when $s = \{i_1, \ldots, i_{n(s)}\}$ are the labels of the units in the sample and z_{i_j} is the observed value of the characteristic of interest for unit i_j. Thus the set of possible sample points depends on both the parameter space and the design. To keep this dependence in mind we will denote the sample space by

$$\begin{aligned} Z(\mathcal{Y}, p) &= \{(s, z_s) : p(s) > 0 \text{ and } z_s = \mathbf{y}(s) \\ &\qquad \text{for some } \mathbf{y} \in \mathcal{Y}\}. \end{aligned} \tag{1.3}$$

If the parameter space is $\mathcal{R}^N$ then the sample space contains uncountably many data points. However for a fixed design and a fixed parameter point there are only finitely many points of the sample space which receive positive probability. Hence this is really

a discrete problem and we denote the probability function by

$$f_{\mathbf{y}}(z) = f_{\mathbf{y}}(s, z_s) = \begin{cases} p(s) & \text{if } z_s = \mathbf{y}(s) \\ 0 & \text{otherwise.} \end{cases} \tag{1.4}$$

The family of possible probability measures defined by this family of probability functions is denoted by $\mathcal{P}$. This notation for finite population sampling is more or less consistent with the standard notation for a statistical inference or decision problem. The only difference is that in finite population sampling the usual practice is to denote the parameter by $\mathbf{y}$, instead of a Greek letter, say θ, as in the other areas of statistics.

In the subsequent chapters we will consider the problem of estimating various functions of the parameter, e.g., the population mean, total, median and variance. Let $\gamma(\mathbf{y})$ denote the function to be estimated. For point estimation problems our loss function will often be squared error, but for now we let $L(\gamma(\mathbf{y}), a)$ denote a general loss when a is the estimate of $\gamma(\mathbf{y})$. Let δ denote a typical estimator of $\gamma(\mathbf{y})$. Then its risk function, or expected loss, is given by

$$\begin{aligned} R(\mathbf{y}, \delta, p) &= \sum_{(s, z_s)} L(\gamma(\mathbf{y}), \delta(s, z_s)) f_{\mathbf{y}}(z) \\ &= \sum_{s} L(\gamma(\mathbf{y}), \delta(s, \mathbf{y}(s)) p(s) \\ &= R(\mathbf{y}; \delta), \end{aligned} \tag{1.5}$$

when there is no ambiguity about the design p. For a further discussion of statistical decision theory see Berger (1985).

1.2 Sufficiency

In statistical decision theory it is the accepted convention to begin with a nonempty set, which is the sample space, equipped with a σ-algebra of subsets, along with a family of possible probability measures indexed by an unknown parameter. For such a model it is of interest to find the minimal sufficient statistic, assuming it exists. Now, in general, for such models a minimal sufficient statistic need not exist. However for discrete models, which includes the finite population sampling model, a minimal sufficient statistic always exists and is easy to find. This was demonstrated in Basu and

Ghosh (1967). In the rest of this section we present this result, which is just our gloss on their work.

Let Z denote the sample space, $\mathcal{B}$ the σ-algebra of subsets and $\mathcal{P}$ a family of probability measures on Z indexed by the parameter $\mathbf{y}$. The triple $(Z,\mathcal{B},\mathcal{P})$ is said to be a **discrete model** if (i) $\mathcal{B}$ is the class of all subsets of Z and (ii) each $P_{\mathbf{y}} \in \mathcal{P}$ is a discrete probability measure. (We are also assuming that for each $z \in Z$ there exist a $P_{\mathbf{y}} \in \mathcal{P}$ such that $P_{\mathbf{y}}(\{z\}) = P_{\mathbf{y}}(z) > 0$.) Note that a discrete model is undominated if and only if Z is uncountable.

Now a statistic is just a function, T, defined on Z. By our choice of $\mathcal{B}$ every such function T is measurable. On the space Z every statistic T defines an equivalence relation $z \sim z'$ when $T(z) = T(z')$. This leads to a partition of Z into equivalence classes of points. Since we need not distinguish between statistics that induce the same partition of Z, we may think of a statistic T as a partition $\Upsilon = \{v\}$ of Z into a family of mutually exclusive and collectively exhaustive parts, v.

Using the usual measure-theoretic definition of sufficiency, Basu and Ghosh prove the following theorem for discrete models. We will give our version of their proof that uses only the pre-measure-theoretic definition of sufficiency.

Theorem 1.1 *If $(Z,\mathcal{B},\mathcal{P})$ is a discrete model, then a necessary and sufficient condition for a statistic (partition) $T = (\Upsilon = \{v\})$ to be sufficient is that there exists a positive real valued function g on Z such that, for all $\mathbf{y} \in \mathcal{Y}$ and $z \in Z$*

$$P_{\mathbf{y}}(z) = g(z)P_{\mathbf{y}}(v_z)$$

where v_z is the member of the partition $\Upsilon = \{v\}$ that contains the point z.

Proof. We will first assume that T is sufficient and show that there exists a g which satisfies the above equation.

Suppose T is sufficient and that $T(z_1) = T(z_2) = t$. Then we will show that $P_{\mathbf{y}}(z_1) > 0$ if and only if $P_{\mathbf{y}}(z_2) > 0$. Suppose not, then there must exist a $\mathbf{y}'$ which gives positive probability to z_1 and zero probability to z_2. But by assumption there exists a $\mathbf{y}''$ which gives positive probability to z_2. But then the conditional probability of z_2 given the event $T = t$ is zero under $\mathbf{y}'$ and strictly positive under $\mathbf{y}''$ which is a contradiction, since T is sufficient. It follows from the above fact that if T is sufficient and $P_{\mathbf{y}}(T = t) > 0$ then $P_{\mathbf{y}}(z) > 0$ whenever $T(z) = t$.

Let $\mathbf{y}$ and t be such that $P_{\mathbf{y}}(T = t) > 0$. If $T(z) \neq t$ then the conditional probability of z given the event $T = t$ is zero. While if $T(z) = t$

$$P_{\mathbf{y}}(z|T = t) = \frac{P_{\mathbf{y}}(z)}{P_{\mathbf{y}}(T = t)} = g(z)$$

which is independent of $\mathbf{y}$ by sufficiency. In either case the equation of the theorem is true for all *yvec* and all z by the above argument and the fact that one side is zero if and only if the other side is zero. Note that $g(z)$ is strictly positive for every z. This follows from the fact that for each z there is at least one $\mathbf{y}$ which assigns it positive probability.

The proof in the other direction is even easier. We suppose the equation holds and then show that T is sufficient. Let z be given and suppose that $\mathbf{y}$ is such that $P_{\mathbf{y}}(\upsilon_z) > 0$. Then

$$P_{\mathbf{y}}(z) = g(z)P_{\mathbf{y}}(\upsilon_z)$$

or when $T(z) = t$

$$P_{\mathbf{y}}(z) = g(z)P_{\mathbf{y}}(T = t)$$

or

$$g(z) = \frac{P_{\mathbf{y}}(z)}{P_{\mathbf{y}}(T = t)} = P_{\mathbf{y}}(z|T = t).$$

If z is such that $T(z) \neq t$ then $P_{\mathbf{y}}(z|T = t) = 0$ and hence T is sufficient and the proof is complete. □

With this theorem one can easily find the minimal sufficient statistic for a discrete model. Recall that the statistic (partition) $T = (\Upsilon = \{\upsilon\})$ is said to be **wider** than the statistic $T^* = (\Upsilon^* = \{\upsilon^*\})$ if every υ is a subset of some υ^*. That is, if every υ^* is a union of a number of the υ's. A statistic (partition) is **minimal sufficient** if it is sufficient and any other sufficient statistic (partition) is wider than it.

With these definitions in mind, we will now define a partition and then show that it is minimal sufficient for the discrete model. For each $z \in Z$ let

$$\mathcal{Y}_z = \{\mathbf{y} : P_{\mathbf{y}}(z) > 0\}.$$

Consider the binary relation defined on Z by $z \sim z'$ if $\mathcal{Y}_z = \mathcal{Y}_{z'}$ and $P_{\mathbf{y}}(z)/P_{\mathbf{y}}(z')$ is a constant in $\mathbf{y}$ for all $\mathbf{y} \in \mathcal{Y}_z = \mathcal{Y}_{z'}$. Clearly this defines an equivalence relationship on Z and hence defines a partition of Z. We denote this partition by $\Upsilon^* = \{\upsilon^*\}$.

We first show that this partition is sufficient. Let z be fixed and $\mathbf{y} \in \mathcal{Y}_z$. Let v_z^* be the member of the partition which contains z. Since for each $z' \in v_z^*$ we have that $P_{\mathbf{y}}(z)/P_{\mathbf{y}}(z')$ is a constant in $\mathbf{y}$ over $\mathcal{Y}_z$ then so is $P_{\mathbf{y}}(z)/P_{\mathbf{y}}(v_z^*)$ constant in $\mathbf{y}$ over $\mathcal{Y}_z$. Let

$$g(z) = P_{\mathbf{y}}(z)/P_{\mathbf{y}}(v_z^*)$$

Then we have

$$P_{\mathbf{y}}(z) = g(z)P_{\mathbf{y}}(v_z^*)$$

for each z and $\mathbf{y} \in \mathcal{Y}_z$. But for $\mathbf{y} \notin \mathcal{Y}_z$ both sides of the above equation are zero, so it is true for all z and $\mathbf{y}$. So by the theorem the partition $\Upsilon^* = \{v^*\}$ is sufficient.

Let $\Upsilon = \{v\}$ be any other sufficient partition. It remains to show that it is wider than the partition $\Upsilon^* = \{v^*\}$. Let z and z' both belong to the same set in the partition Υ, say v. Then by remarks made in the proof of the theorem $P_{\mathbf{y}}(z) > 0$ if and only if $P_{\mathbf{y}}(z') > 0$ and in this case it follows from the theorem that

$$P_{\mathbf{y}}(z)/P_{\mathbf{y}}(z') = g(z)/g(z')$$

which is a constant independent of $\mathbf{y}$. Hence z and z' belong to the same set in the partition Υ and the partition Υ is wider than the partition Υ^*.

The minimal sufficient statistic has an alternative characterization. For each $z \in Z$ let $L_z(\mathbf{y})$ be the likelihood function, i.e.

$$L_z(\mathbf{y}) = \begin{cases} P_{\mathbf{y}}(z) & \text{for } \mathbf{y} \in \mathcal{Y}_z \\ 0 & \text{for } \mathbf{y} \notin \mathcal{Y}_z \end{cases} \tag{1.6}$$

and

$$\bar{L}_z(\mathbf{y}) = L_z(\mathbf{y}) / \sup_{\mathbf{y}} L_z(\mathbf{y}) \tag{1.7}$$

be the standardized likelihood function. Consider the mapping $z \to \bar{L}_z(\cdot)$, a mapping of Z into a class of real-valued functions on $\mathcal{Y}$. It is easy to see that if $\bar{L}_z(\mathbf{y}) = \bar{L}_{z'}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{Y}$, then z and z' belong to the same set in the partition Υ. That is, this mapping is a minimal sufficient statistic, and induces the minimal sufficient partition given above.

In the next section we will identify the minimal sufficient statistic in finite population sampling.

1.3 The sufficiency and likelihood principles

The sufficiency and likelihood principles were widely used in other areas of statistics before their role in finite population sampling was properly understood. Here we summarize some of the arguments of Basu (1969) which shows how they are applied in finite population sampling. For more detail see Ghosh (1988) which is a collection of Basu's work in this area.

The sufficiency principle states that if T is a sufficient statistic and $T(z) = T(z')$ then the inference about $\mathbf{y}$ should be the same whether the sample is z or z'. This principle has gained wide acceptance. In discrete models since the mapping $z \to \bar{L}_z(\mathbf{y})$ is a minimal sufficient statistic, according to the sufficiency principle two sample points are equally informative if

$$\bar{L}_z(\mathbf{y}) = \bar{L}_{z'}(\mathbf{y}) \text{ for all } \mathbf{y}.$$

Note that the sufficiency principle does not say anything about the nature of the information supplied by z. For this we need the likelihood principle which states that all the information contained in z is embodied in the standardized likelihood function $\bar{L}_z$. To see the implications for finite population sampling let us return to the notation of Section 1.1.

Suppose the parameter space $\mathcal{Y}$ and design p are fixed and yield the sample space $Z(\mathcal{Y}, p)$. Then for a fixed sample point $z = (s, z_s)$

$$\begin{aligned} \mathcal{Y}_z = \mathcal{Y}_{(s,z_s)} &= \{\mathbf{y} : f_{\mathbf{y}}(z) > 0\} \\ &= \{\mathbf{y} : \mathbf{y}(s) = z_s\}. \end{aligned}$$

From this we easily see that the standardized likelihood function is given by

$$\begin{aligned} \bar{L}_z(\mathbf{y}) = \bar{L}_{(s,z_s)}(\mathbf{y}) &= 1 \quad \text{if } \mathbf{y} \in \mathcal{Y}_z \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{1.8}$$

Because the mapping $z \to \bar{L}_z(\mathbf{y})$ is a minimal sufficient statistic and the likelihood function is constant over $\mathcal{Y}_z$, all we learn from the observed data $z = (s, z_s)$ are the values of the characteristic of interest for the units in the sample and that the 'true' $\mathbf{y}$ must be with consistent with these observed values. Note that this is a very simple and intuitively appealing notion and agrees with many people's naive idea about what is learned from the observed sample. Note that this observation is independent of the sampling design. That is, after the sample $z = (s, z_s)$ is observed the minimal suffi-

cient partition does not depend, in any way, on the value of $p(s)$. (In fact, Basu demonstrated that this is true even for sequential sampling plans where, at any stage, the choice of the next population unit to be observed is allowed to depend only on the observed values of the characteristic of the previously selected units.)

For many frequentist statisticians, this last observation is perhaps the most bothersome aspect of the above argument. That is, that the likelihood principle implies that the design probabilities should not be considered in analysing the data, after the sample has been observed. In particular, choosing an estimator which is unbiased for a given design violates the likelihood principle. But from one point of view this is not surprising when one recalls the way probability arises in finite population sampling. Since the characteristic y_i is assumed to be measured without error the only way probability enters into the model is through the design p. Hence the phenomenon of randomness is not inherent within the problem but is inserted into the problem by the statistician. In other areas of statistics the statistician uses probability theory to model uncontrollable randomness, while in finite population sampling the whole analysis is based on controlled randomness introduced by the statistician.

Godambe (1966) had noted before Basu that the application of the likelihood principle to finite population sampling would mean that the sampling design is irrelevant for data analysis. But he, as many other non-Bayesian statisticians since then, has chosen to ignore the likelihood principle and tried to justify a role for the design when analysing the data. For two sophisticated attempts see Scott (1977) and Sugden and Smith (1984). They considered situations where some information available to the person who designed the survey is not available to the one who must analyse the data. They argued that in such situations the design may become informative. Although such examples are interesting, we do not feel that they lessen the force of the above argument.

Recall that the likelihood principle in finite population sampling justifies a very natural notion; that is, given the observed data $z = (s, z_s)$ one just learns the values of z_s and that the unsampled y_j's for $j \notin s$ must come from a $\mathbf{y}$ which is consistent with z_s. So the basic question of finite population sampling is how can one relate the unseen, $\mathbf{y}(s') = \{ y_j : j \notin s \}$, to the seen, z_s? Without some assumptions about how these two sets are related, knowing z_s does not tell one anything at all about $\mathbf{y}(s')$. Presumably, for a

frequentist, the design p along with the unbiasedness requirement is a way to relate the unseen to the seen. We, however, have never found these arguments compelling.

On the other hand, the Bayesian paradigm allows one to relate the unseen to the seen in a straightforward way which does not violate the likelihood principle. Let $\pi(\mathbf{y})$ denote the prior density or probability function of the Bayesian statistician. The statistician chooses $\pi(\cdot)$ to represent and summarize his or her prior beliefs and information about $\mathbf{y}$. Given the sample $z = (s, z_s)$ one then computes the conditional density of $\mathbf{y}$ given z, say $\pi(\mathbf{y}|z)$. This is concentrated on the set $\mathcal{Y}_z$ and is just π with the seen, z_s, inserted in their appropriate places and renormalized, so it integrates to one over $\mathcal{Y}_z$. Then for squared error loss the Bayes estimator against π for the population total is

$$\sum_{i \in s} z_i + \sum_{j \notin s} E_\pi(y_j|z), \tag{1.9}$$

where for $j \notin s$, $E_\pi(y_j|z)$ is the conditional expectation of y_j with respect to $\pi(\mathbf{y}|z)$. The form of the Bayes estimator emphasizes that estimation in finite population sampling can be thought of as a prediction problem, i.e. of predicting the unseen from the seen and one should argue conditionally from the seen to the unseen.

One of the things that first got us interested in finite population sampling was trying to understand why Bayesian ideas were not used more in the area. In the first place, finite population sampling is the one area of statistics in which everyone seems to agree that useful prior information is available and should be used. Moreover there is an elegant argument, the one given above, that suggests a Bayesian approach is a sensible one. However, most practitioners proceed in a way which is not only not Bayesian but seems to be antithetical to the Bayesian paradigm. Furthermore there were not any arguments giving a noninformative Bayesian justification for standard frequentist methods, as is often the case in other areas of statistics.

The main goal of this monograph is to argue that Bayesian ideas should play a role in both the theory and practice of finite population sampling. We will show how the Bayesian approach can be applied in a number of different contexts with a variety of types of prior information. In particular we will give a noninformative Bayesian justification for many of the standard frequentist meth-

ods and show how this reasoning can lead to sensible procedures in problems which the usual theory finds quite difficult. Although our orientation is definitely Bayesian, we are still interested in the frequentist properties of the procedures we present and will discuss them in some detail.

1.4 A Bayesian example

In many statistical inference problems it can be quite difficult to specify a sensible prior distribution and then carry out a Bayesian analysis. This is particularly true for problems with a large dimensional parameter space, of which finite population sampling forms an important subset. For such large-scaled problems a Bayesian analysis seems impossible without some simplifying assumptions that allow us to model our prior information. A variety of such models are needed to be able to handle various amounts of prior knowledge. We now present some results of Ericson (1969b) which are applicable when reasonable prior guesses for the population mean and variance are present.

The simplest and most basic of the sampling designs is simple random sampling without replacement. This is usually used when little is known about the population and, in particular, the labels carry scant information about the units. We will discuss this situation in great detail in Chapter 2. Now, however, we will consider the case where, although the labels contain no information, we do have a sensible prior guess for the mean of the population, say m. How should we choose a prior distribution to reflect this prior information?

A naive first guess might be to assume that the y_i's are independent and identically distributed (or iid in what follows) with common mean m. Under this prior, we see from (1.9), that the Bayes estimator of the population total, under squared error loss, given $z = (s, z_s)$ is $\sum_{i \in s} z_i + (N - n(s))m$. Because of the independence we have that the seen, z_s, gives us no information about the unseen, $\mathbf{y}(s') = \{ y_j : j \notin s \}$. In order to relate the unseen to the seen we need a prior which makes the y_i's dependent but still carries no information about the characteristic in the labels. Such a prior for the y_i's would be invariant or symmetric under permutations of the labels and hence exchangeable. We now give an example of a technique for defining exchangeable distributions which is quite useful in finite population sampling.

Let θ be a real valued parameter. We will assume that θ has a probability distribution given by the probability density function h. Moreover we assume that given θ the y_i's are iid with probability density function $g(\cdot|\theta)$ Unconditionally, i.e integrating out θ, this defines a probability density function for $\mathbf{y}$ given by

$$\pi(\mathbf{y}) = \int_{-\infty}^{\infty} \prod_{i=1}^{N} g(y_i|\theta)h(\theta)\, d\theta. \tag{1.10}$$

Here θ is sometimes called a hyperparameter and is introduced as a mixing parameter to generate exchangeable distributions from independent identically distributed distributions. This is somewhat related to the model-based approach to finite population sampling (Royall, 1970). In both the model based and the Bayesian approaches $\mathbf{y}$ has a probability distribution. In addition both emphasize prediction and the design probabilities play no role in the inferences after the sample has been observed. However, the model-based approach assumes a family of possible distributions for $\mathbf{y}$ which is indexed by some parameter which is then estimated by frequentist methods.

Let $z = (s, z_s)$ be the observed sample point. Then given the sample the posterior density is zero for each $\mathbf{y} \notin \mathcal{Y}_z$. But for $\mathbf{y} \in \mathcal{Y}_z$ we have that

$$\begin{aligned}
\pi(\mathbf{y}|z) &= \int_{-\infty}^{\infty} \prod_{j \notin s} g(y_j|\theta) \prod_{i \in s} g(y_i|\theta)h(\theta)\, d\theta \\
&\quad \div \int \cdots \int \left\{ \int_{-\infty}^{\infty} \prod_{j \notin s} g(y_j|\theta) \prod_{i \in s} g(y_i|\theta)h(\theta)\, d\theta \right\} d\prod_{j \notin s} y_j \\
&= \int_{-\infty}^{\infty} \prod_{j \notin s} g(y_j|\theta) \frac{\prod_{i \in s} g(y_i|\theta)h(\theta)}{\pi(y_i;\, i \in s)}\, d\theta \\
&= \int_{-\infty}^{\infty} \prod_{j \notin s} g(y_j|\theta)h(\theta|z)\, d\theta,
\end{aligned} \tag{1.11}$$

where $\pi(y_i;\, i \in s)$ is the joint marginal density function of the units in the sample and, following a standard abuse of notation, $h(\theta|z)$ is the conditional density of θ given the sample z, under the model defined above. That is, $\{ y_j : j \notin s \}$ are iid $g(\cdot|\theta)$ given θ and θ has density $h(\cdot|z)$.

We now consider an important special case of the above setup. Given θ we assume that the y_i's are iid each normally distributed with mean θ and variance σ^2, i.e. each normal(θ, σ^2). We assume that θ is normal (m, τ^2) where m, σ^2 and τ^2 are all assumed to

be known. Note that these three parameters are easily interpreted. The parameter m is your prior guess for the mean of the population, τ^2 is a measure of how certain you are about your choice of m and σ^2 is your prior guess for the amount of variability in the population. As we have just seen this defines a prior distribution over $\mathbf{y}$, which in turn induces a prior distribution for the population mean, $\sum_{i=1}^{N} y_i/N$. Our first step is to find this distribution.

Note that the previous model has the following equivalent formulation. Let $w_1, \ldots, w_N$ be iid each normal $(0, \sigma^2)$ and independent of the distribution for θ, then we have that $y_i = w_i + \theta$. Clearly the y_i's are jointly normally distributed with $E(y_i) = m$, $\mathrm{Var}(y_i) = \sigma^2 + \tau^2$ and $\mathrm{Cov}(y_i, y_j) = \tau^2$. Using these facts one easily finds that under this prior distribution the distribution of $\sum_{i=1}^{N} y_i/N$ is normal $(m, \tau^2 + \sigma^2/N)$.

Suppose now we have observed the data $z = (s, z_s)$ containing $n(s)$ units. Then the sample mean is denoted by

$$\bar{z} = \bar{z}_s = \sum_{i:\, i \in s} z_i/n(s). \tag{1.12}$$

For the rest of this section we will let $n(s) = n$. To compute the posterior distribution of the unseen given the seen, we see from (1.11) that we need to find $h(\theta|z)$. But this is easy since conditional on θ the seen, $z(s) = \{ z_i : i \in s \}$, are iid each normal$(\theta,\sigma^2)$ and our prior for θ is normal (m, τ^2). From this it follows that $h(\theta|z)$ is a normal (m', v') density function where

$$m' = \frac{\tau^2 \bar{z}_s + m(\sigma^2/n)}{\tau^2 + \sigma^2/n}$$

and

$$v' = \frac{\sigma^2/n}{\tau^2 + \sigma^2/n}\tau^2.$$

From this fact, (1.11) and using the same argument that led to the prior distribution of the population mean, we see that given the sample, the posterior distribution for $\mathbf{y}(s') = \{ y_j : j \notin s \}$, the unseen, is multivariate normal with common mean $= m'$, common variance $= \sigma^2 + v'$ and common covariance $= v'$. With these facts in hand it is easy to find the posterior expectation of the population mean and its posterior variance as well.

For a fixed sample s we denote the population mean by

$$\mu = \mu(y) = N^{-1}\Big(\sum_{i:\, i\in s} y_i + \sum_{j:\, j\notin s} y_j\Big). \tag{1.13}$$

From this it follows that

$$E(\mu|z) = N^{-1}\big(n\bar{z}_s + (N-n)m'\big) \tag{1.14}$$

and

$$\begin{aligned}
\mathrm{Var}(\mu|z) &= \frac{1}{N^2}\big\{(N-n)(\sigma^2+v') + (N-n)(N-n-1)v'\big\} \\
&= \frac{N-n}{N^2}\big\{\sigma^2 + (N-n)v'\big\} \\
&= \frac{N-n}{N^2}\left\{\sigma^2 + (N-n)\frac{(\sigma^2/n)\tau^2}{\tau^2+\sigma^2/n}\right\} \\
&= \frac{N-n}{N}\left\{\frac{\sigma^2/n}{\tau^2+\sigma^2/n}(\tau^2+\sigma^2/N)\right\} \\
&= \frac{N-n}{N}\frac{\sigma^2/n}{\tau^2+\sigma^2/n}\mathrm{Var}(\mu).
\end{aligned} \tag{1.15}$$

Note that both the estimator and its posterior variance have nice forms with the values m, $\sigma^2$and τ^2 entering in sensible ways. As we saw before the Bayes estimator replaces each unseen value with its posterior expectation, which by exchangeability is the value m' for each of the unobserved units. m' is a convex mixture of the sample mean, $\bar{z}_s$ and m, our prior guess for the population mean. Their relative weights depend on σ^2 and τ^2 in a natural way, which is quite common in such Bayesian models. The posterior variance consists of three factors. The first is just the usual finite population correction factor which appears in many variance formulas in finite population sampling. The second is just the ratio of the posterior variance of θ to the prior variance of θ. Both these factors are less than one and when multiplied times the third factor, the prior variance of μ, give the posterior variance of μ.

This is a simple but appealing model. It demonstrates how certain kinds of prior information can be used in finite population sampling in a straightforward Bayesian manner. But both its simplicity and lack of flexibility limit its usefulness in studying real populations. Ericson (1969b) also considered more general and more realistic versions of this model. Even this simple model, however, includes ideas that are quite useful in other Bayesian approaches.

The notions of mixing iid distributions to get an exchangeable prior and of hierarchical normal modeling will play a large role in what follows.

1.5 Posterior linearity

The results of the previous section can be generalized to distributions other than the normal. For example, if conditional on θ, $y_1, \ldots, y_N$ are iid having a common pdf belonging to the one-parameter exponential family with the form

$$f(y|\theta) = \exp(\theta y - \psi(\theta))h(y), \tag{1.16}$$

and θ has the conjugate prior pdf of the form

$$\pi(\theta) \propto \exp(\alpha\theta - \nu\psi(\theta)), \tag{1.17}$$

then the posterior pdf of θ given $z_i, i \in s$ is

$$\pi(\theta|z_i, i \in s) \propto \exp\left[\left(\sum_{i\in s} z_i + \alpha\right)\theta - (n+\nu)\psi(\theta)\right]. \tag{1.18}$$

It is well known that $E(y_i|\theta) = \psi'(\theta)$and $V(y_i|\theta) = \psi''(\theta)/n$. Also, it is easy to verify after integration by parts that(see e.g. Raiffa and Schlaiffer (1961))

$$E[\psi'(\theta)|z_i, i \in s] = \left(\sum_{i\in s} z_i + \alpha\right)/(n+\nu) = (n\bar{z}_s + \alpha)/(n+\nu), \tag{1.19}$$

where $\bar{z}_s = n^{-1}\sum_{i\in s} z_i$. From the prior distribution of θ given in (1.17), it can be verified after integration by parts that

$$E[\psi'(\theta)] = \alpha/\nu; \quad E[\psi''(\theta)] = \nu V[\psi'(\theta)]. \tag{1.20}$$

Thus it is possible to express the posterior mean of $\psi'(\theta)$ alternately as

$$\frac{\bar{z}_s V(\psi'(\theta)) + (\alpha/\nu)(\nu/n)V[\psi'(\theta)]}{V(\psi'(\theta)) + (\nu/n)V(\psi'(\theta))} = \frac{\bar{z}_s V(\psi'(\theta) + E(\psi'(\theta))EV(\bar{Y}|\theta)}{V(\psi'(\theta)) + EV(\bar{Y}|\theta)}. \tag{1.21}$$

The above result was obtained by Ericson (1969b) under less restrictive conditions. Indeed, Ericson did not invoke any distribu-

tional assumptions, but needed instead an assumption of 'posterior linearity' which will be discussed below.

Suppose that conditional on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$, a vector of unknown parameters, the y_i have a common mean $\mu(\boldsymbol{\theta})$, and $\boldsymbol{\theta}$ has a prior under which

$$E[\mu(\boldsymbol{\theta})|z] = \alpha \bar{z}_s + \beta, \tag{1.22}$$

where α and β do not depend on the y_i. Then (1.22) is referred to as the condition of 'posterior linearity'.

The notion of posterior linearity was introduced independently by Ericson (1969b), and Hartigan (1969), and was exploited very effectively later in a series of papers by Goldstein (see e.g. Goldstein (1975)). Diaconis and Ylvisaker (1979) pointed out that if conditional on an unknown parameter θ, the y_i are iid with a pdf given in (1.16), then the prior pdf of θ must be of the conjugate form given in (1.17).

The following theorem was proved in Ericson (1969b).

Theorem 1.2 *Suppose that conditional on an unknown parameter* $\boldsymbol{\theta}$, y_i *have a common mean* $\mu(\boldsymbol{\theta})$,*and (1.22) holds. Suppose, in addition* $V(y_i|\boldsymbol{\theta}) < \infty$ *for all* i, *and* $0 < V(\mu(\boldsymbol{\theta})) < \infty$. *Then*

$$E[\mu(\boldsymbol{\theta})|z] = [\tau^2 \bar{z}_s + mE(V(\bar{z}_s|\boldsymbol{\theta}))]/[\tau^2 + E(V(\bar{z}_s|\boldsymbol{\theta}))], \tag{1.23}$$

where $E(\mu(\boldsymbol{\theta})) = m$, *and* $V(\mu(\boldsymbol{\theta})) = \tau^2$.

Proof. First note that $E(\bar{z}_s) = EE(\bar{z}_s|\boldsymbol{\theta}) = E\mu(\boldsymbol{\theta}) = m$. Hence,

$$m = EE(\mu(\boldsymbol{\theta})|z) = E(\alpha \bar{z}_s + \beta) = \alpha m + \beta \tag{1.24}$$

so that $\beta = (1 - \alpha)m$. Now observe that

$$E[\bar{z}_s \mu(\boldsymbol{\theta})] = E[\mu(\boldsymbol{\theta}) E(\bar{z}_s|\boldsymbol{\theta})] = E[\mu^2(\boldsymbol{\theta})] = V(\mu(\boldsymbol{\theta})) + m^2, \tag{1.25}$$

and

$$\begin{aligned} E[\bar{z}_s \mu(\boldsymbol{\theta})] &= E[\bar{z}_s E\mu(\boldsymbol{\theta})|z] \\ &= E[\alpha \bar{z}_s^2 + (1-\alpha) m \bar{z}_s] \\ &= \alpha[V(\bar{z}_s + m^2] + (1-\alpha)m^2 \\ &= \alpha V(\bar{z}_s) + m^2. \end{aligned} \tag{1.26}$$

Noting that $V(\bar{z}_s) = EV(\bar{z}_s|\boldsymbol{\theta}) + VE(\bar{z}_s|\boldsymbol{\theta})$ the theorem follows immediately from (1.22) and (1.24)–(1.26). □

Remark. In the special case of a one-parameter exponential family with conjugate prior, (1.23) reduces to (1.21). We emphasize though that Ericson's model includes distributions in addition to

the ones belonging to the one-parameter exponential family. An important example is the multivariate-t distribution. Also, if the y_i are iid with a distribution function F, and F has a Dirichlet process prior, then posterior linearity still holds. The present model of Ericson will be used later in Chapter 4 in studying the robustness of certain empirical Bayes procedures.

The present result of Ericson can be immediately extended to finite population sampling. Under Ericson's model, for the finite population mean μ, the Bayes estimator is

$$E[\mu|z] = N^{-1}[n\bar{z}_s + (N-n)E(\mu(\boldsymbol{\theta})|z)]. \tag{1.27}$$

We shall refer to (1.27) repeatedly in later chapters.

1.6 Overview

We conclude this chapter with an overview of the rest of the book. We assume that the reader is familiar with the basic fundamentals of the frequentist approach to finite population sampling as found in Cochran (1977) or Särndal *et al.* (1992). We also assume that the reader is familiar with the basic ideas of statistical inference as found in a first-year graduate course in statistical theory.

This is not meant to be an encyclopedic treatise of Bayesian methods in finite population sampling. Most of what follows is based on our research and reflects our interests. The theme of Chapters 2 and 3 is how one can incorporate different types of prior information into a finite population problem without actually specifying a prior distribution. Furthermore one's inferential procedures are implemented in the usual Bayesian manner and are based on a pseudo posterior distribution. This pseudo posterior reflects the prior information and the sample data however and the design plays no role in the final inferences. Chapter 4 discusses an empirical Bayes approach to some problems in finite population sampling while Chapter 5 discusses a corresponding hierarchical Bayes approach to these problems.

In Chapter 2 we give a noninformative Bayesian justification of the basic frequentist inferential procedures in finite population sampling. This justification is based on the Polya posterior which is a predictive distribution for the unobserved units in the population given the observations in the sample. It is appropriate when the statistician believes the observed sample is representative of the population as a whole, i.e. in the judgement of the statisti-

cian the 'seen' and 'unseen' units of the population are roughly exchangeable. The Polya posterior does not depend on the design probabilities and can be used just like a ordinary posterior to make inferences even though it does not arise from any single prior. The Polya posterior has in fact a stepwise Bayes justification which means that inferential procedures based on it are admissible. The Polya posterior can be used for both point and set estimation problems for a variety of parameters of interest although in most cases the procedures cannot be found explicitly but must be approximated by simulation.

In Chapter 3 we consider some extensions of the Polya posterior which allow one to incorporate different kinds of prior information into the analysis. In each case there is a stepwise Bayes justification for the resulting pseudo posteriors. We first consider the situations where there is either a prior guess for the entire population or a prior guess for each member of the population. Next we consider the situation where an auxiliary variable is available and the statistician's beliefs about the ratios of the characteristic of interest to the auxiliary variable are exchangeable. Then we study stratified populations and consider two situations with different levels of prior knowledge about the stratification. We next consider the problem of nonresponse. We show that the Polya posterior along with an assumed model for the relationship between the responders and nonresponders leads to methods similar in spirit to those arising from multiple imputation. Next we consider some nonparametric problems and show that there is a close relationship between admissibility questions for these problems and admissibility questions in finite population sampling. Finally we consider one situation where the necessary beliefs about about exchangeability are not present. We assume that the units in the population are labelled in such a way that members of the population whose labels are close together are more alike than members whose labels are far apart. For such a case an estimator that linearly interpolates between successive members of the sample seems intuitively sensible. This estimator is shown to be stepwise Bayes and the underlying pseudo posterior is shown to lead to sensible interval estimators.

The final two chapters of this monograph consider applications of empirical and hierarchical Bayes methods in finite population sampling. These methods, which have been gaining increasing popularity in recent years, are particularly well-suited to tackle small area estimation problems. The term 'small area' is commonly used

to denote a small geographical area, such as a county, municipality or a census division. It may also describe a 'small domain', that is a small subpopulation or group of people within a large geographical area with a specified age, sex and race. Sample survey data usually provide reliable estimators of totals and means for large areas or domains. However, the usual direct survey estimators for a small area, based on data only from the sample units in the area, are likely to yield unacceptably large standard errors due to an unduly small sample size in that area. Sample sizes for small areas are typically small because the overall sample size in a survey is usually determined to provide accuracy at a much higher level of aggregation than that of small areas. Thus, there is a need to connect the different small areas either explicitly or implicitly through a model. Both empirical and hierarchical Bayes methods do this modelling very effectively by building exchangeability among the different small areas. Both these approaches recognize the uncertainty in specifying accurately the prior parameters. The difference between the two approaches is that whereas an empirical Bayes method estimates the unknown prior parameters through the marginal distributions of the observations, a hierarchical Bayes method models the uncertainty by putting a suitable prior (often diffuse) to the prior parameters.

In the chapter on empirical Bayes methods, we first show how some of the stepwise Bayes estimators developed earlier can also be viewed as empirical Bayes estimators. Next, such estimators are derived for simultaneous estimation of several small area means based on a normal model. The normality assumption is next relaxed, and is replaced by the assumption of posterior linearity. Next in this chapter, similar estimators are derived for small area means based on two-stage surveys. The results are then extended to include situations where area-specific auxiliary information is available. Finally, in this chapter, estimators are derived when there is unit-specific rather than area-specific auxiliary information.

In the chapter on hierarchical Bayes methods, first a synthesis between model- and design-based estimators is achieved via a hierarchical Bayes approach. The Horvitz–Thompson estimator which is well known as a design-based estimator is viewed also as a model-based estimator. It is also shown that without employing some prior information, design-based estimators can often lead to meaningless answers. Next, estimates of small area means are derived in the absence of any covariates. The results are then ex-

tended in the presence of area-specific covariates. In a special case, these estimators are identified as best linear unbiased predictors, and also as best unbiased predictors under an added normality assumption of the errors. Similar results are obtained in the presence of unit-specific covariates. Finally, in this chapter, small area estimates are derived using a generalized linear model.

CHAPTER 2

A noninformative Bayesian approach

In many areas of statistics it is possible to give a noninformative Bayesian justification for standard frequentist methods. However, until recently such a justification was not available in finite population sampling. In the first chapter we summarized the Bayesian and predictive approach to finite population sampling and noted that in this approach the design probabilities play no role in the final analysis. We believe that it was this fact which delayed such a noninformative Bayesian justification in finite population sampling. In this chapter we develop such a justification. The theoretical argument underlying this justification is based on admissibility. The first real admissibility work in finite population sampling was done by V. M. Joshi in the mid-1960s. His method of proof was in the spirit of admissibility arguments current in other areas of statistics at that time. Unfortunately these arguments were quite technical and seemed to us too complicated and somehow not in keeping with the spirit of finite population sampling. Here we will prove the admissibility of a variety of point estimators in finite population sampling using a different type of argument. Our method of proof will use the stepwise Bayes technique. The stepwise Bayes technique was first given in Johnson (1971) and named in Hsuan (1979). In Section 2.1, to demonstrate the stepwise Bayes technique, we will prove the admissibility of the maximum likelihood estimator in a binomial problem. In Section 2.2 we will prove a complete class theorem for problems with a finite sample space and finite parameter space. Using these ideas we then prove in Section 2.3 the admissibility of the sample mean when estimating the population mean in a finite population sampling problem. In Section 2.4 we show how the stepwise Bayes technique can yield admissible set estimation procedures for certain problems in finite population sampling. We shall see that even though there is no single underlying prior distribution against which the sample mean

is Bayes, given any data point there is always a pseudo 'posterior distribution' which one can use in the usual Bayesian way and which yields the sample mean as a pseudo Bayes procedure. We will then identify this pseudo 'posterior distribution' with a Polya urn distribution and name it the 'Polya posterior'. We then argue that it leads to sensible point and interval estimators for a variety of problems. In particular it will yield a noninformative Bayesian justification for some of the standard frequentist methods. In Section 2.5 we briefly recall some facts about the Polya urn distribution. In Section 2.6 we show how this distribution is related to the stepwise Bayes argument of Section 2.3.2. In Section 2.7 we show how point and set estimators based on the Polya posterior can be found approximately in practice. In Section 2.8 we compute point and interval estimators based on the Polya posterior for a variety of problems. We shall see that it often yields procedures similar to those of standard frequentist theory. In addition it can be used in situations where the standard methods are difficult to apply.

2.1 A binomial example

The stepwise Bayes method of proving admissibility was introduced in Johnson (1971) although a similar argument was given earlier in Wald and Wolfowitz (1951). In this section we give an example of the technique in a binomial problem. This problem was also considered by Johnson. We consider this example, not only because it is a nice introduction to the method, but the generalization from the binomial setting to the multinomial setting is very basic; underlies much that follows.

Let Z denote the sample space of a binomial(n, θ) random variable where n is known and $\theta \in [0,1]$ is unknown. Consider the problem of estimating θ with squared error as the loss function. Let δ denote an estimator, then its risk function is given by

$$R(\theta, \delta) = \sum_{z=0}^{n} (\delta(z) - \theta)^2 f_\theta(z)$$

where $f_\theta(z) = \binom{n}{z}\theta^z(1-\theta)^{n-z}$ is the usual binomial probability function.

Let π be a prior probability distribution for θ, then the Bayes

risk of an estimator, δ, is given by

$$r(\delta, \pi) = \int_0^1 R(\theta, \delta)\, d\pi(\theta).$$

δ_π is Bayes against π if $r(\delta_\pi, \pi) = \inf_\delta r(\delta, \pi)$. For this problem it is well known that for a given π, which assigns positive marginal probability to each member of the sample space, the unique δ_π is just the conditional expectation of θ given z, whenever the resulting Bayes risk is finite.

Suppose now that π is the beta(α, β) distribution. That is, its density function on the interval $(0, 1)$ is given by

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where $\alpha > 0$ and $\beta > 0$ and $\Gamma(\cdot)$ is the usual gamma function. Recall that $E(\theta) = \alpha/(\alpha+\beta)$ and $\mathrm{Var}(\theta) = \alpha\beta/\{(\alpha+\beta)^2(\alpha+\beta+1)\}$. For this π, $\pi(\theta|z) \propto \theta^{\alpha+z-1}(1-\theta)^{\beta+n-z-1}$ and so the posterior distribution of $\theta|z$ is just Beta$(\alpha+z, \beta+n-z)$ and $\delta_\pi(z) = (z+\alpha)/(n+\alpha+\beta)$. Note as $\alpha \to 0$ and $\beta \to 0$ this estimator approaches $\delta^{mle}(z) = z/n$, the maximum likelihood estimator of θ. This suggests that the δ^{mle}, might be 'Bayes' against the improper prior $1/\{\theta(1-\theta)\}$. This cannot be right, since under this improper prior, $E(\theta|z)$ is undefined when $z = 0$ and $z = n$ (i.e. the integrals do not exist), but for $z = 1, \ldots, n-1$ all the integrals do exist and yield the value z/n.

It has long been known that there is a close relationship between admissibility and being Bayes. The above suggests that the maximum likelihood estimator, $\delta^{mle}(z) = z/n$, is almost Bayes in some sense and one might hope that this fact could be used to prove it is admissible. This is in fact what the stepwise Bayes technique does. It demonstrates the admissibility of δ^{mle} by showing that δ^{mle} is Bayes against a sequence of two priors, i.e. it becomes 'Bayes' in two steps. Before we begin we need one more bit of notation. For a prior distribution π let $q(z; \pi) = \int_0^1 f_\theta(z)\, d\pi(\theta)$.

The first prior, π_1, puts mass 0.5 at 0 and mass 0.5 at 1. Then $q(z; \pi_1) = 0.5$ when z is 0 or n and is 0 when $z = 1, \ldots, n-1$. If θ is the probability that a certain coin comes up heads, then π_1 asserts that you believe that the coin either has two tails or two heads and each of these two possibilities is equally likely. Under such a model one expects to see either all heads or all tails with equal probability and no other outcomes are possible. Note that

under π_1 we have that $\pi_1(0|z=0) = \pi_1(1|z=n) = 1$. Moreover $\pi_1(\cdot|z)$ is undefined for $z = 1, \ldots, n-1$. Hence, any estimator which estimates 0 when $z = 0$ and 1 when $z = n$ is Bayes against π_1. So, δ^{mle} is Bayes against π_1 along with many other estimators. Now some of the estimators in this class of Bayes estimators against π_1 are admissible while others will be inadmissible. The problem is how to identify which are which.

We do this by considering a restricted problem, which is essentially the original problem with the sample points and parameter points already taken care of, removed from consideration. That is, our sample space is the set $\{1, \ldots, n-1\}$ and our parameter space is (0, 1). For this restricted problem the probability function is just the original probability function renormalized so that it sums to one over the new and smaller sample space. Clearly it will be given by

$$f_\theta^*(z) = f_\theta(z)/\{1 - \theta^n - (1-\theta)^n\}$$

for $z \in \{1, \ldots, n-1\}$ and $\theta \in (0, 1)$.

We now have a well-defined restricted problem and for this new problem we can consider the question of estimating θ with squared error loss. We will approach this problem in the Bayesian manner and pick a second prior, say π_2, and compute the corresponding Bayes estimator. Our choice for this second prior will be

$$\begin{aligned} \pi_2(\theta) &\propto \{1 - (1-\theta)^n - \theta^n\}/\{\theta(1-\theta)\} \\ &\propto \sum_{k=1}^{n-1} \binom{n}{k} \theta^{k-1}(1-\theta)^{n-1-k} \end{aligned}$$

which is a bounded function of θ and so π_2 can be made into a density function by the proper choice of a normalizing constant. Then for the restricted problem $\pi_2(\theta|z) \propto f_\theta^*(z)\pi_2(\theta) \propto \theta^{z-1}(1-\theta)^{n-z-1}$ which is beta$(z, n-z)$. Hence for $z = 1, \ldots, n-1$ we have that $E(\theta|z) = z/n$ is the unique Bayes estimator of θ in the restricted problem. Putting these two parts together we see that δ^{mle} is the unique stepwise Bayes estimator of θ for the original problem.

Although a proper definition of stepwise Bayes will not be given until the next section, it essentially formalizes the above procedure. You begin by selecting some prior distribution. If the resulting Bayes estimator is uniquely defined for every point in the sample space, you are done and the estimator is admissible. On the other hand if the Bayes estimator is not defined at some points in the

sample space you construct the restricted problem and select a second prior and find its Bayes estimator. If it is uniquely defined for every point in the second sample space, you are done and the resulting estimator is admissible. On the other hand if there are still sample points in the restricted problem where the Bayes estimator is not defined, you construct a new restricted problem and repeat the process again. You continue in this way until you have an estimator which is defined at every sample point. The resulting estimator is said to be stepwise Bayes. Typically this will ensure that the estimator is admissible as well. In the next section we will show that this is true for problems with a finite sample space and finite parameter space. We shall now show that the fact that δ^{mle} is stepwise Bayes in the above problem ensures that it is admissible.

Suppose that δ^{mle} is not admissible, then there exists an estimator, say δ, which dominates it, i.e.

$$R(\theta, \delta) \leq R(\theta, \delta^{mle}) \text{ for all } \theta \tag{2.1}$$

and strict for at least one $\theta \in [0,1]$. But $R(0, \delta^{mle}) = 0$ implies that $\delta(0) = 0$ and $R(1, \delta^{mle}) = 0$ implies that $\delta(n) = n$. It follows from this and (2.1) that for $\theta \in (0,1)$ we have

$$\sum_{z=1}^{n-1} (\delta(z) - \theta)^2 f_\theta(z) \leq \sum_{z=1}^{n-1} (\delta^{mle}(z) - \theta)^2 f_\theta(z)$$

or equivalently, that

$$\sum_{z=1}^{n-1} (\delta(z) - \theta)^2 f^*_\theta(z) \leq \sum_{z=1}^{n-1} (\delta^{mle}(z) - \theta)^2 f^*_\theta(z).$$

Now multipling both sides of this last equation by $\pi_2(\theta)$ and integrating over θ we have that

$$r^*(\delta, \pi_2) \leq r^*(\delta^{mle}, \pi_2)$$

where r^* denotes the Bayes risk for the restricted problem. Hence δ is Bayes against π_2 for the restricted problem. But δ^{mle}, restricted to the set $\{1, \ldots, n-1\}$, is the unique Bayes estimator against π_2 for this problem and so $\delta(z) = z/n$ for $z = 1, \ldots, n-1$ from which it follows that $\delta(z) = \delta^{mle}(z)$ for all z which is a contradiction.

Note that we selected the priors π_1 and π_2 with the goal, of proving the admissibility of δ^{mle}, in mind. Actually there are other choices for π_1 which would work as well. Any prior which concentrates all its mass on 0 and 1 would do the job. However, the

choice of π_2 was crucial and in some sense justifies the improper prior $1/\{\theta(1-\theta)\}$ discussed above, since the factor in the numerator in π_2 cancelled with the factor in the denominator of f_θ^*. It is useful to think of the process geometrically. The first step was to select a prior which concentrated its mass on the two endpoints of the parameter space and then in the second step we selected the correct prior on the interior to get the maximum likelihood estimator as stepwise Bayes.

This process generalizes in a straightforward way to proving the admissibility of the maximum likelihood estimator when we have a sample of size, say n, from a multinomial distribution. To see how this would work we consider the trinomial case. Let the three possible values be b_0, b_1 and b_2 with corresponding probabilities $1-\theta_1-\theta_2$, θ_1 and θ_2. The parameter space is the two-dimensional simplex where $\theta_1 \geq 0$, $\theta_2 \geq 0$ and $\theta_1 + \theta_2 \leq 1$. The argument proceeds as follow. In the first stage the prior concentrates all its mass on the three corners. The amount of mass at each corner does not matter as long as each corner gets some positive amount of mass. This takes care of all the sample points where we just see one b_i every time. In the second stage there are three priors each of which concentrates all of its mass on an edge of the simplex. For example on the edge where $\theta_2 = 0$ we put the prior $1/\{\theta_1(1-\theta_1)\}$. Technically this is not correct because we are ignoring the factor which goes in the numerator which makes it a proper prior. But when we compute our stepwise Bayes estimator this factor cancels with the denominator of the renormalized probability function, just as in the binomial case and so we can effectively forget about it. For the restricted problem this prior gives positive marginal probability to all those sample points which have at least one b_0, at least one b_1 and no b_2's. We do a similar thing on the edge $\theta_1 = 0$ to take care of all the sample points which consist of just b_0's and b_2's. On the edge $\theta_1 + \theta_2 = 1$ we essentially put the prior $1/\{\theta_1(1-\theta_1)\}$ or equivalently the prior $1/\{\theta_2(1-\theta_2)\}$ which takes care of all those sample points where both b_1 and b_2 appear but b_0 does not. The order in which these three priors are chosen is immaterial as long as they follow the first stage and precede the third and final stage. Now in the last stage we still have to take care of all the sample points where each of the three b_i's appear at least once. This is done by essentially putting the prior $1/\{(1-\theta_1-\theta_2)\theta_1\theta_2\}$ on the interior of the simplex. Again we are ignoring the factor in the numerator because it cancels when we do the computations. It is

easy to check that this sequence of priors does indeed yield the maximum likelihood estimator as its stepwise Bayes estimator.

The generalization to an arbitrary multinomial problem should now be clear. Again the maximum likelihood estimator is an admissible stepwise Bayes estimator. At the first stage mass must be concentrated on the corners. The second stage considers the one-dimensional edges in turn. The third stage considers all the two-dimensional faces in turn and so on until in the final stage our prior is concentrated on the interior of the simplex. For additional details see Brown (1981). This stepwise Bayes argument is important for, as we shall see, it underlies many of the admissibility results that follow.

2.2 A characterization of admissibility

In this section we will give a minimal complete class theorem, i.e. we will identify the class of admissible estimators for problems with finite sample space and finite parameter space. Since we will be applying this theorem to problems in finite population sampling we will attempt to keep the notation consistent with that of Section 1.1. In particular the parameter will be denoted by y instead of θ. The parameter y may or may not be vector valued; this is immaterial in what follows.

Let Z denote the sample space which contains finitely many values. The σ-algebra of measurable sets is the class of all subsets of Z. Let $\{ f_y : y \in \mathcal{Y} \}$ be a family of probability functions defined on Z, where $\mathcal{Y} = \{ y_1, \ldots, y_k \}$. The elements of $\mathcal{Y}$ may be real numbers or vectors of real numbers. Their actual form is not important; what is important, however, is that $\mathcal{Y}$ contains just finitely many elements. Assume that for each $z \in Z$, $f_{y_i}(z) > 0$ for at least one $y_i \in \mathcal{Y}$. Consider the problem of estimating some real valued function of y, say $\gamma(y)$. The decision space is $\mathcal{A}$, a closed and bounded interval of real numbers. The nonnegative loss incurred in estimating $\gamma(y)$ with a an element of $\mathcal{A}$ is $L(\gamma(y), a)$. We assume for each fixed $\gamma(y)$, $L(\gamma(y), \cdot)$ is a strictly convex function on $\mathcal{A}$. Let $\pi = (\pi_1, \ldots, \pi_k)$ denote a prior distribution on $\mathcal{Y}$. The marginal probability function on Z under π is given by $q(z; \pi) = \sum_{i=1}^{k} f_{y_i}(z)\pi_i$. For any estimator δ of $\gamma(y)$ its Bayes risk

is given by

$$r(\delta,\pi) = \sum_i \sum_z L(\gamma(y_i),\delta(z)) f_{y_i}(z)\pi_i$$
$$= \sum_{z:q(z;\pi)>0} \sum_i \{L(\gamma(y_i),\delta(z)) f_{y_i}(z)\pi_i / q(z;\pi)\} q(z;\pi). \tag{2.2}$$

Thus if δ_π is an estimator which is Bayes against π, it is unique on the set of z's for which $q(z;\pi) > 0$. Such an estimator need not be admissible, although a unique Bayes estimator must be admissible. On the other hand, since $\mathcal{Y}$ is finite, an admissible estimator must be Bayes with respect to some prior distribution (Ferguson, 1967).

In the case when $f_{y_i}(z) > 0$ for each z and each y_i, $q(z;\pi) > 0$ for each z and for every prior π. Then, an estimator is admissible if and only if it is unique Bayes with respect to some prior π. In this section we consider the more general case where for some choices of y_i and z, $f_{y_i}(z)$ may be zero. Before we state our theorem we need one more bit of notation. For the prior distribution π let $\mathcal{Y}(\pi) = \{y_i : \pi_i > 0\}$. For a pair of prior distributions, π and π' let $(\pi,\pi') = \sum_{i=1}^k \pi_i\pi_i'$. Note that $(\pi,\pi') = 0$ if and only if $\mathcal{Y}(\pi) \bigcap \mathcal{Y}(\pi')$ is the empty set.

The following theorem was stated without proof in Meeden and Ghosh (1981a) because it is just a restatement of a result of Hsuan (1979).

Theorem 2.1 *If δ is an admissible estimator, then there exists a non-empty set of prior distributions, $\pi^1 = (\pi_1^1,\ldots,\pi_k^1)^T,\ldots,$ $\pi^m = (\pi_1^m,\ldots,\pi_k^m)^T$ such that*
(i) $(\pi^i,\pi^j) = 0$ for $i \neq j$;
(ii) if

$$\Lambda^1 = \{z : q(z;\pi^1) > 0\}$$

and for $i = 2,\ldots,m$

$$\Lambda^i = \left\{ z : q(z;\pi^i) > 0 \text{ and } z \notin \bigcup_{j=1}^{i-1} \Lambda^j \right\},$$

then each Λ^i is non-empty and $\bigcup_{i=1}^m \Lambda^i = Z$;
(iii) if δ_{π^i} denotes a Bayes estimator of $\gamma(y)$ against π^i, then for $i = 1,\ldots,m$

$$\delta(z) = \delta_{\pi^i}(z) \text{ for all } z \in \Lambda^i. \tag{2.3}$$

Conversely, if δ is an estimator which satisfies conditions (i), (ii) and (iii) for some sequence of prior distributions, $\pi^1, \ldots, \pi^m$, then δ is admissible.

Proof. Let δ be an admissible estimator, we will find a sequence of priors which satisfies the three conditions of the theorem. Since δ is admissible and the parameter space $\mathcal{Y}$ is finite there exists a prior distribution, say π^1, against which δ is Bayes. The set $\Lambda^1 = \{ z : q(z;\pi^1) > 0 \}$ is non-empty. On this set δ_{π^1} is unique and δ satisfies part (iii) for $i = 1$. If $\Lambda^1 = Z$ we are done. So suppose $Z - \Lambda^1$ is non-empty. Then $\mathcal{Y} - \mathcal{Y}(\pi^1)$ must also be non-empty.

Note that for each $z \in Z - \Lambda^1$, there exists at least one $y \in \mathcal{Y} - \mathcal{Y}(\pi^1)$ such that $f_y(z) > 0$. Let

$$\mathcal{Y}_o(\pi^1, \Lambda^1) = \{ y : y \in \mathcal{Y} - \mathcal{Y}(\pi^1),\ f_y(z) = 0 \text{ for all } z \in Z - \Lambda^1 \}.$$

Now $\mathcal{Y} - \mathcal{Y}(\pi^1) - \mathcal{Y}_o(\pi^1, \Lambda^1)$ is also non-empty and for any y in this set

$$c(y) = \sum_{z \in Z - \Lambda^1} f_y(z) > 0.$$

Consider now the restricted problem where $z \in Z - \Lambda^1$ and $y \in \mathcal{Y} - \mathcal{Y}(\pi^1) - \mathcal{Y}_o(\pi^1, \Lambda^1)$. For this restricted problem, the family of possible probability functions is

$$\{ f_y^* : y \in \mathcal{Y} - \mathcal{Y}(\pi^1) - \mathcal{Y}_o(\pi^1, \Lambda^1) \}$$

where for $z \in Z - \Lambda^1$

$$f_y^*(z) = f_y(z)/c(y).$$

We now let $\bar{\delta}$ denote the restriction of δ to $Z - \Lambda^1$. If $\bar{\delta}$ is inadmissible for the restricted problem, then there exists some estimator $\bar{\delta}_o$ which dominates it. Let

$$\delta_o(z) = \begin{cases} \delta(z) & \text{for } z \in \Lambda^1 \\ \bar{\delta}_o(z) & \text{for } z \in Z - \Lambda^1. \end{cases}$$

Then, for the original problem , δ_o dominates δ, since for $y \in \mathcal{Y}(\pi^1) \cup \mathcal{Y}_o(\pi^1, \Lambda^1)$ and $z \in Z - \Lambda^1$, $f_y(z) = 0$. Since this is a contradiction, $\bar{\delta}$ must be admissible for the restricted problem. Hence, for the restricted problem, $\bar{\delta}$ is Bayes with respect to some prior distribution which concentrates its mass on the set $\mathcal{Y} - \mathcal{Y}(\pi^1) - \mathcal{Y}_o(\pi^1, \Lambda^1)$. For each y_i in this set let $\alpha(y_i)$ be the mass assigned

to y_i. Let $\pi^2 = (\pi_1^2, \ldots, \pi_k^2)$ be the prior which is defined by

$$\pi_i^2 = \begin{cases} 0 & \text{for } y_i \in \mathcal{Y}(\pi^1) \cup \mathcal{Y}_o(\pi_1, \Lambda^1) \\ d\alpha(y_i)c(y_i) & \text{for } y_i \in \mathcal{Y} - \mathcal{Y}(\pi^1) - \mathcal{Y}_o(\pi^1, \Lambda^1) \end{cases}$$

where d is a constant chosen so that $\sum_{i=1}^k \pi_i^2 = 1$. Note that $(\pi^1, \pi^2) = 0$, Λ^2 is non-empty and (iii) is satisfied for $z \in \Lambda^2$. If $\Lambda^1 \cup \Lambda^2 = Z$ we are done. If not, we consider the restricted problem with z restricted to the set $Z - \Lambda^1 - \Lambda^2$ and proceed as in the above. Continuing in this fashion we see that if δ is admissible (i), (ii) and (iii) of the theorem must be satisfied for some finite set of prior distributions.

Conversely, now suppose δ is an estimator which satisfies (i), (ii) and (iii) of the theorem. If δ is not admissible then there exists some estimator δ^* such that

$$E_y L(\gamma(y), \delta^*) \leq E_y L(\gamma(y), \delta) \text{ for all } y \in \mathcal{Y} \tag{2.4}$$

with strict inequality for some $y \in \mathcal{Y}$. Now consider the restricted problem with z restricted to Λ^1 and y restricted to $\mathcal{Y}(\pi^1)$. In view of (iii), δ restricted to Λ^1 is the unique Bayes estimator for this problem. From this, we have $\delta^*(z) = \delta(z)$ for $z \in \Lambda^1$ and equality holds in (2.4) for $y \in \mathcal{Y}(\pi^1)$, since for $y \in \mathcal{Y}(\pi^1)$ and $z \notin \Lambda^1$, $f_y(z) = 0$. Now consider the restricted problem with z restricted to Λ^2 and y restricted to $\mathcal{Y}(\pi^2)$. Just as before we have that $\delta^*(z) = \delta(z)$ for $z \in \Lambda^2$ and equality holds in (2.4) for $y \in \mathcal{Y}(\pi^2)$. Proceeding in this way we see that (2.4) implies that $\delta^*(z) = \delta(z)$ for all z and so δ is admissible. □

As of yet, we have not given a formal definition of stepwise Bayes. For the setting of this section we will say that an estimator, δ, is **stepwise Bayes** if there exists a finite sequence of priors such that the three conditions of the theorem are satisfied. Hence for such problems the theorem just states that an estimator is admissible if and only if it is stepwise Bayes. Throughout the book we will demonstrate the admissibility of a variety of estimators from finite population sampling. In all the cases we will first prove the admissibility for a parameter space of the type, $\mathcal{Y}(\mathbf{b})$, see (1.1), by using the theorem of this section. Then the admissibility for the parameter space, $\mathcal{R}^N$, will follow easily.

2.3 Admissibility of the sample mean

In this section we will prove the admissibility of the sample mean for estimating the population mean under squared error loss, using a stepwise Bayes argument. This result was originally proved in Joshi (1965) by a different method. In Section 1.3 we saw how the likelihood principle applied to finite population sampling yielded the fact that the design probabilities should play no inferential role after the data has been observed. This suggests that the admissibility of an estimator should essentially be independent of the design. This was in fact demonstrated in Scott (1975) and for this reason the design will play little role in the proofs that follow. Before proving the result for the sample mean we present the result of Scott.

2.3.1 Admissibility and the design

We will be using the notation of Section 1.1. Let δ be an estimator for estimating $\gamma(\mathbf{y})$ with loss function $L(\cdot,\cdot)$. The parameter space is $\mathcal{Y}$ and p is a design. We say that the design p is **absolutely continuous** with respect to the design p_o (written $p \ll p_o$) if $p_o(s) = 0 \Rightarrow p(s) = 0$. That is, p is absolutely continuous with respect to p_o if every sample that can be observed under p can also be observed under p_o. The following theorem states that if δ is an admissible estimator under the design p_o and if p is absolutely continuous with respect to p_o then δ is also admissible under p. The assumption of absolute continuity is a very weak one and is necessary just to guarantee that the estimator is well defined for all possible samples under the second design.

Theorem 2.2 *If δ is admissible under p_o and $p \ll p_o$ then δ is admissible under p.*

Proof. Let $S_o = \{ s : p_o(s) > 0 \}$ and $w = \min_{s \in S_o} p_o(s)/p(s)$. Note that $w \leq 1$ and $w\,p(s)/p_o(s) \leq 1$ for all $s \in S_o$. Let $v(s) = w\,p(s)/p_o(s)$, so that $0 \leq v(s) \leq 1$ for all $s \in S_o$. Also $p(s) = 0 \Rightarrow v(s) = 0$.

If δ is not admissible under p then there exists an estimator δ^* which dominates it for the design p. We now define a new randomized estimator, δ', by

$$\delta'(s, z_s) = \begin{cases} \delta^*(s, z_s) & \text{with probability} \quad v(s) \\ \delta(s, z_s) & \text{with probability} \quad 1 - v(s) \end{cases}$$

and show that it dominates δ under the design, p_o, which is a contradiction.

Now for a fixed $\mathbf{y}$ we have

$$\begin{aligned} R(\mathbf{y},\delta',p_o) &= \sum_{s\in S_o} p_o(s)\{v(s)L(\gamma(\mathbf{y}),\delta^*(s,z_s)) \\ &\quad + (1-v(s))L(\gamma(\mathbf{y}),\delta(s,z_s))\} \\ &= w\sum_{s\in S_o} p(s)L(\gamma(\mathbf{y}),\delta^*) + \sum_{s\in S_o} p_o(s)(1-v(s))L(\gamma(\mathbf{y}),\delta) \\ &\leq w\sum_{s\in S_o} p(s)L(\gamma(\mathbf{y}),\delta) + \sum_{s\in S_o} p_o(s)(1-v(s))L(\gamma(\mathbf{y}),\delta) \\ &= \sum_{s\in S_o} p_o(s)v(s)L(\gamma(\mathbf{y}),\delta) + \sum_{s\in S_o} p_o(s)(1-v(s))L(\gamma(\mathbf{y}),\delta) \\ &= \sum_{s\in S_o} p_o(s)L(\gamma(\mathbf{y}),\delta) \\ &= R(\mathbf{y},\delta,p_o) \end{aligned}$$

where the step with the inequality follows from the assumption that δ^* dominates δ under p and the two surrounding equalities follow from the definition of $v(s)$. Since the inequality holds for all $\mathbf{y}$ in the parameter space and is strict for at least one, the proof is complete. □

Note that this result is quite general; there are no restrictions on the function to be estimated, the loss function and the design.

2.3.2 The proof

As before we let $\mu = \mu(\mathbf{y})$ denote the population mean. Here we consider the problem of estimating the population mean with squared error loss. Let δ^{sm} denote the estimator which is just the sample mean, i.e.

$$\delta^{sm}(z) = \bar{z} = \bar{z}_s = \textstyle\sum_{i\in s} z_i/n(s). \tag{2.5}$$

The admissibility of the sample mean will be demonstrated for two different parameter spaces.

Theorem 2.3 *Let* $\mathbf{b} = (b_1,\ldots,b_k)^T$ *be a vector of distinct real numbers and let* p *be some design. Then for estimating the population mean with squared error loss the sample mean,* δ^{sm}*, is admissible, under any design* p*, when the parameter space is* $\mathcal{Y}(\mathbf{b})$.

Proof. The admissibility of δ^{sm} will be demonstrated by showing that it is a stepwise Bayes estimator and applying Theorem 2.1. If π is a prior and $z = (s, z_s)$ is the observed data it follows from (1.9) that the Bayes estimate of μ is

$$\delta_\pi(z) = N^{-1}\left\{\sum_{i\in s} z_i + \sum_{j\notin s} E_\pi(y_j|z)\right\}. \tag{2.6}$$

Now if π is such that for a given z and for each $j \notin s$ we have $E_\pi(y_j|z) = \bar{z} = \delta^{sm}(z)$ then $\delta_\pi(z) = \bar{z} = \delta^{sm}(z)$. So if we can produce a finite sequence of priors such that at each step, and for every z with positive probability at the given step, the conditional expectation of each unobserved unit given the observed data is just the sample mean the admissibility of δ^{sm} will be proved. The way the argument will proceed is that first we will take care of all data points where all the observed units take on the same value. In the second stage we will take care of all data points where all the observed units just take on two different values. In the next stage where they just take on three different values and so on, until all the possibilities have been accounted for.

Before we begin we need some more notation. For a $\mathbf{y} \in \mathcal{Y}(\mathbf{b})$ and $i = 1, \ldots, k$ let

$$c_y(i) = \text{number of } y_j\text{'s which equal } b_i,$$

and for a sample s

$$c_y(i, s) = \text{number of } y_j\text{'s in } y(s) \text{ which equal } b_i,$$

and finally for a data point $z = (s, z_s)$

$$c_z(i, s) = \text{number of } z_j\text{'s in } z_s \text{ which equal } b_i.$$

In the first stage we take as the prior π^1 the distribution which puts mass $1/k$ on the k points $(b_1, \ldots, b_1)^T, \ldots, (b_k, \ldots, b_k)^T$ and zero mass elsewhere. Under this prior the only points in our sample space, $Z(\mathcal{Y}(\mathbf{b}), p)$, which receive positive marginal probability are those where all the observed values are identical. Let z be such a data point, where the common value is, say, b_i. For such an observed data point it is easy to check that for an unobserved unit j

$$P_{\pi^1}(y_j = b_i|z) = 1$$

From this it follows that $E_{\pi^1}(\mu|z) = b_i$ which is just the sample mean.

In the second stage we will take care of all those data points where just two distinct values appear. First we handle all those points where just b_1 and b_2 appear. Let

$$\mathcal{Y}^*(b_1, b_2) = \mathcal{Y}(b_1, b_2) - \{(b_1, \ldots, b_1)^T \cup (b_2, \ldots, b_2)^T\}.$$

Then $\mathcal{Y}^*(b_1, b_2)$ is the parameter space for the restricted problem, and on it we define our second prior to be

$$\begin{aligned} \pi^2(\mathbf{y}) &\propto \int_0^1 \theta^{c_y(1)-1}(1-\theta)^{c_y(2)-1}\, d\theta \\ &= \{\Gamma(c_y(1))\Gamma(c_y(2))\}/\Gamma(N). \end{aligned}$$

For a probability function of this form it is easy to find the marginal probability of any subset. In particular, for a fixed $\mathbf{y}$ and s with $c_y(1, s) \geq 1$ and $c_y(2, s) \geq 1$

$$\begin{aligned} \pi^2(\mathbf{y}(s)) &\propto \sum_{\mathbf{y}':\mathbf{y}'(s)=\mathbf{y}(s)} \pi^2((y_1', \ldots, y_N')^T) \\ &= \int_0^1 \left\{ \sum_{\mathbf{y}':\mathbf{y}'(s)=\mathbf{y}(s)} \theta^{c_{y'}(1)-1}(1-\theta)^{c_{y'}(2)-1} \right\} d\theta \\ &= \int_0^1 \theta^{c_y(1,s)-1}(1-\theta)^{c_y(2,s)-1}\, d\theta \\ &= \{\Gamma(c_y(1, s))\Gamma(c_y(2, s))\}/\Gamma(n(s)). \end{aligned}$$

Using this fact it is easy to calculate the conditional distribution of any unobserved unit given an observed data point $z = (s, z_s)$ which just contains the values b_1 and b_2. For $j \notin s$ we have that

$$\begin{aligned} P(y_j = b_1 | z) &= P(y_j = b_1 \text{ and } \mathbf{y}(s) = z_s)/P(\mathbf{y}(s) = z_s) \\ &= \frac{\Gamma(c_z(1, s) + 1)\Gamma(c_z(2, s))}{\Gamma(n(s) + 1)} \frac{\Gamma(n(s))}{\Gamma(c_z(1, s))\Gamma(c_z(2, s))} \\ &= c_z(1, s)/n(s). \end{aligned}$$

Hence for such a z and $j \notin s$

$$\begin{aligned} E(y_j | z) &= b_1\{c_z(1, s)/n(s)\} + b_2\{c_z(2, s)/n(s)\} \\ &= \bar{z}_s. \end{aligned}$$

Next we consider $\mathcal{Y}^*(b_1, b_3)$ and so on until we have accounted for all data points where just a pair of values appear.

In the next stage we will take care of all the data points where exactly three distinct values appear in the observed units. Let

$\mathcal{Y}^*(b_1, b_2, b_3)$ consist of all those vectors where b_1, b_2 and b_3 all appear at least once and no other values appear. This will be the parameter space for the restricted problem at this step. On this set we take as our prior

$$\begin{aligned} \pi_l(\mathbf{y}) &\propto \int_0^1 \int_0^1 \theta_1^{c_y(1)-1} \theta_2^{c_y(2)-1} (1-\theta_1-\theta_2)^{c_y(3)-1} \, d\theta_1 d\theta_2 \\ &= \{\Gamma(c_y(1))\Gamma(c_y(2))\Gamma(c_y(3))\}/\Gamma(N). \end{aligned} \tag{2.7}$$

Just as before the marginal probability of any subset is easily found. It follows, for a data point z with just these three values appearing, that for $j \notin s$

$$P(y_j = b_i|z) = c_z(i,s)/n(s) \tag{2.8}$$

for $i = 1, 2$ and 3. From this we have that $E(y_j|z)$ is the sample mean for every unobserved unit. So at this step, under this prior, the Bayes estimator is just the sample mean.

In the next stage we will account for data points where four different values appear and so on, until we have taken care of all possible values. When we actually stop will depend on k, the length of $\mathbf{b}$, and the samples s which receive positive probability under p. This completes the proof of the theorem, since we have demonstrated that the sample mean is indeed a stepwise Bayes estimator and hence by Theorem 2.1 is admissible. □

With this result it is easy to prove the admissibility of the sample mean for the usual parameter space.

Theorem 2.4 *For estimating the population mean, under squared error loss, the sample mean, δ^{sm}, is admissible, under any design p, when the parameter space is $\mathcal{R}^N$.*

Proof. Suppose δ^{sm} is not admissible, then there exists some estimator, say δ, which dominates it, i.e.

$$R(\mathbf{y}, \delta, p) \leq R(\mathbf{y}, \delta^{sm}, p)$$

for all $\mathbf{y} \in \mathcal{R}^N$ with strict inequality for at least one point, say $\mathbf{y}_o$. Let $\mathbf{b}$ be the vector of distinct values that occur in $\mathbf{y}_o$. Now the previous equation must hold for all $\mathbf{y} \in \mathcal{Y}(\mathbf{b})$, with strict inequality at $\mathbf{y}_o$. This implies that δ^{sm} is inadmissible when the parameter space is $\mathcal{Y}(\mathbf{b})$, which is a contradiction. Hence δ^{sm} must be admissible when the parameter space is $\mathcal{R}^N$. □

This theorem was first proved in Joshi (1965) by a different argument. The previous proof was first given in Meeden and Ghosh

(1983). Note that in the proof of Theorem 2.3 the actual values of the b_i's play no role whatsoever. It is this fact that allows for the same proof to work for every choice of $\mathbf{b}$. But once we have the admissibility of δ^{sm} for every choice of $\mathbf{b}$ then the admissibility for $\mathcal{R}^N$ follows at once. Formally we say that an estimator is **finitely admissible** if given any point in the parameter space there exists a finite subset of the parameter space which contains the given point and for which the estimator is admissible when the parameter is restricted to the finite subset. Clearly finite admissibility implies admissibility but the converse need not be true. Hence we have shown that the sample mean is finitely admissible when estimating the population mean by considering $\mathcal{Y}(\mathbf{b})$ as a possible parameter space. This finite parameter space was introduced in Hartley and Rao (1968), in what they called the scale load situation.

The two theorems can easily be generalized in two important directions. First of all squared error loss can be replaced by any strictly convex loss function. Secondly, the population mean can be replaced by any function of the parameter of interest. Using the same sequence of priors, the above proof yields the admissibility of the resulting stepwise Bayes estimator. Now in most of these other cases this estimator cannot be found in closed form. However, it can be found approximately through simulation. As we shall see, these admissible estimators often have good frequentist properties and yield sensible solutions for problems where the standard frequentist methods are quite difficult to apply. In the following we will consider several such examples.

In Section 1.3 we argued that the basic question in finite population sampling was how to relate the unseen to the seen, without violating the likelihood principle, in particular, without using the design probabilities. As we saw, the Bayesian paradigm was one way to accomplish this. However it is not clear how a straightforward Bayesian argument can justify the sample mean as an estimator of the population mean, in situations with little prior information. Note that the preceding argument does in fact give a noninformative Bayesian justification for the sample mean in such situations. This follows because one need not specify a prior distribution. But given any possible data point, z, the stepwise Bayes argument yields a 'posterior distribution' of an unseen unit y_j, given z, which is just the empirical distribution function of the sample. This is exactly the content of (2.8) and seems to reflect how many

people view the unseen, after seeing the observed data, in problems with little or no prior information. Also note that at each stage of the argument the prior is exchangeable and arises from a Dirichlet-like mixture of a multinomial model. In any case, we believe that the results of this section suggest the possibility that given a data point, z, the resulting 'posterior distribution', which arises from the stepwise Bayes argument, should lead to a sensible noninformative Bayesian analysis. This idea introduces one of the major themes of this book and one which will be discussed extensively in later chapters. That is, how various levels of prior information can be incorporated into finite population sampling problems in a Bayesian and pseudo Bayesian manner.

2.4 Set estimation

Practitioners often seem to prefer a set estimate to a point estimate. For a frequentist this is a confidence interval while for a Bayesian it is a credible set, usually a highest posterior density region. In Meeden and Vardeman (1985) it was demonstrated how admissible Bayesian credible sets can be constructed. In this section we show how this approach can be modified to the stepwise Bayes situation to yield admissible set estimators for the setup described in Section 2.2. Unfortunately such admissible credible sets are somewhat different from highest posterior probability regions and are often difficult to find in practice. In Section 2.7 we discuss a compromise approach we will use to find sensible set estimates for problems when a stepwise Bayes approach is being employed.

We now show how a sequence of priors which led to an admissible point estimator can yield an admissible set estimator as well. We will use the notation of Section 2.2. Let Z be the sample space which contains finitely many values. Let $\mathcal{Y}$ denote the parameter space which also contains finitely many values. For each $y \in \mathcal{Y}$ let $f_y(\cdot)$ be the probability function on Z when y is the true state of nature. Let $\gamma(y)$ be the function for which a set estimator is desired and $\Gamma = \{\gamma(y) : y \in \mathcal{Y}\}$. A nonrandomized confidence procedure is a set $SE \subseteq \Gamma \times Z$. If a statistician is using SE and z is observed then

$$SE(z) = \{\gamma : (\gamma, z) \in SE\}$$

is the set estimate of $\gamma(y)$.

The appropriateness of a confidence procedure depends on at

least two of its characteristics, namely the size of the subsets of Γ it produces and the probabilities with which those subsets fail to include a true but unknown value of $\gamma(y)$. So we define

$$n_{SE}(y) = \sum_z n(SE(z)) f_y(z)$$

and

$$p_{SE}(y) = \sum_z I[\gamma(y) \notin SE(z)] f_y(z)$$

where $n(\cdot)$ denotes the number of elements in the set and where I denotes the indicator function of a set. Since one hopes to simultaneously control $n_{SE}(y)$ and $p_{SE}(y)$, we say that SE_1 is **admissible** provided there is no other procedure SE_2 with

$$n_{SE_2}(y) \leq n_{SE_1}(y) \quad \forall\, y$$

and

$$p_{SE_2}(y) \leq p_{SE_1}(y) \quad \forall\, y$$

where for some y at least one of the two inequalities is strict.

Let π be a prior probability distribution over $\mathcal{Y}$. For a confidence procedure SE let

$$n_{SE}(\pi) = \sum_y n_{SE}(y)\pi(y)$$

and

$$p_{SE}(\pi) = \sum_y p_{SE}(y)\pi(y).$$

Let $w \in (0, 1/2)$. We say that SE is w-**Bayes** versus π provided there is no confidence procedure SE_1 with $w n_{SE_1}(\pi) + (1 - w) p_{SE_1}(\pi) < w n_{SE}(\pi) + (1 - w) p_{SE}(\pi)$. In this formulation of the set estimation problem, a Bayesian must not only specify a prior distribution but also must decide on the relative importance of set size and noncoverage probability, which is reflected in the choice of w. The larger the choice of w the more importance there is attached to announcing a small set and less importance is given to stating a small noncoverage probability.

Suppose now $\pi(y) > 0\ \forall\, y \in \mathcal{Y}$ and let $q(z; \pi) = \sum_y f_y(z)\pi(y)$, the marginal probability function on Z under π. Then the posterior probability function of $\gamma(y)$ given a value z is

$$\pi(\gamma | z) = \sum_{\{y : \gamma(y) = \gamma\}} \frac{f_y(z)\pi(y)}{q(z; \pi)}.$$

It follows from Proposition 4.1 of Meeden and Vardeman (1985) that any confidence procedure SE satisfying

$$\{ (\gamma, z) \ : \ \pi(\gamma|z) > w/(1-w) \} \subset SE$$

and

$$\{ (\gamma, z) \ : \ \pi(\gamma|z) < w/(1-w) \} \subset SE'$$

is w-Bayes against the prior distribution π and is admissible, where SE' denotes the complement of SE. Note that for a given w a w-Bayes procedure is not unique if there exists at least one z and one γ for which $\pi(\gamma|z) = w/(1-w)$. This lack of uniqueness means that for the set estimation problem, a general minimal complete class result for stepwise Bayes procedures analogous to Theorem 2.1, for the point estimation problem, cannot be proved following the earlier argument where the uniqueness of the procedures played a crucial role. However, it is easy to show that certain special cases of stepwise Bayes set procedures are admissible.

Let $\pi^1, \ldots, \pi^m$ be a sequence of mutually orthogonal prior distributions on $\mathcal{Y}$. Following the notation of Theorem 2.1 we let

$$\Lambda^1 = \{ z \ : \ q(z : \pi^1) > 0 \}$$

and for $i = 2, \ldots, m$,

$$\Lambda^i = \{ z \ : \ q(z : \pi^i) > 0 \text{ and } z \notin \cup_{j=1}^{i-1} \Lambda^j \}.$$

We assume that each Λ^i is non-empty and $\cup_{i=1}^m \Lambda^i = Z$. For $z \in \Lambda^i$ let

$$SE^\star(z) = \{ \gamma \ : \ \pi^i(\gamma|z) > w/(1-w) \}. \tag{2.9}$$

Note that for some z it is possible that $SE^\star(z)$ is the empty set. We now show that the set estimation procedure $SE^\star$ is admissible.

Theorem 2.5 *Let $w \in (0, 1/2)$ be given and if $SE^\star$ is the set estimation procedure defined by (2.9) then $SE^\star$ is admissible.*

Proof. Suppose $SE^\star$ is not admissible then there exists some procedure SE_1 such that

$$n_{SE_1}(y) \leq n_{SE^\star}(y) \quad \forall\, y \tag{2.10}$$

and

$$p_{SE_1}(y) \leq p_{SE^\star}(y) \quad \forall\, y \tag{2.11}$$

where for some y at least one of the two inequalities is strict. We will show that this is not possible since (2.10) and (2.11) imply that SE_1 and $SE^\star$ are identical.

The first step is to show that $SE_1(z) = SE^\star(z)$ for $z \in \Lambda^1$. By construction $SE^\star$ is w-Bayes against the prior distribution π^1. From (2.10) and (2.11) we see that SE_1 is w-Bayes against the prior distribution π^1 as well. But among all procedures SE which are w-Bayes against the prior distribution π^1, $SE^\star$ minimizes $n_{SE}(y)$ uniformly for y in the support of π^1. It follows from (2.10) that $SE_1(z) = SE^\star(z)$ for $z \in \Lambda^1$.

The next step is to show that $SE_1(z) = SE^\star(z)$ for $z \in \Lambda^2$. To do this we consider the restricted problem with sample space $Z - \Lambda^1$ and parameter space $\mathcal{Y} - \mathcal{Y}(\pi^1) - \mathcal{Y}_o(\pi^1, \Lambda^1)$, defined in the proof of Theorem 2.1. For this restricted problem the probability function is

$$f_y^*(z) = f_y(z)/c(y)$$

for $z \in Z - \Lambda^1$ and

$$c(y) = \sum_{y \in Z - \Lambda^1} f_y(z) > 0$$

for each y in the parameter space for the restricted problem. For this restricted problem consider the prior distribution given by $d\pi^2(y)c(y)$ where d is chosen so that this distribution does indeed sum to one over the parameter space for the restricted problem. It is easy to check that the posterior distribution for this restricted problem agrees with $\pi^2(\cdot|z)$ for $z \in \Lambda^2$ and that $SE^\star$ is w-Bayes for this restricted problem as well. In addition, for the restricted problem, among all procedures w-Bayes against $d\pi^2(y)c(y)$ it uniformly minimizes set size for y in the support of π^2. From this and the fact that $SE_1(z) = SE^\star(z)$ for $z \in \Lambda^1$ it follows that $SE_1(z) = SE^\star(z)$ for $z \in \Lambda^2$.

Continuing in this way we show that SE_1 and $SE^\star$ agree on each Λ^i for $i = 1, \ldots, m$ which completes the proof. □

We note that if in the definition of $SE^\star$ we replace the strict inequality with greater than or equal to, the resulting procedure would also be admissible. The proof is the same as that given above except at each stage $SE^\star$ uniformly minimizes the noncoverage probability.

2.5 The Polya urn

In Section 2.3 the admissibility of the sample mean was demonstrated using the stepwise Bayes technique. We saw that even

though there is no single underlying prior distribution given any data point there is always a pseudo 'posterior distribution' which yields the sample mean as the estimate in the usual Bayesian manner. We will now identify this pseudo 'posterior distribution' with a Polya urn distribution and name it the 'Polya posterior'. In this section we will review the Polya urn distribution. A good reference is Feller (1968).

Suppose an urn contains n_1 balls of colour b_1, and n_2 balls of colour b_2. A ball is drawn at random. It is replaced in the urn along with an additional ball of the same colour. A new random drawing is made from the urn, now containing $n_1 + n_2 + 1$ balls and this process is repeated, say, R times. The probability that in these R drawings the first r_1 of them are b_1 and the last r_2 of them are b_2, where $r_1 + r_2 = R$ is given by

$$\frac{n_1(n_1+1)\cdots(n_1+r_1-1)n_2(n_2+1)\cdots(n_2+r_2-1)}{(n_1+n_2)(n_1+n_2+1)\cdots(n_1+n_2+R-1)}.$$

Now consider the probability of any other ordering of r_1 of the b_1's and r_2 of the b_2's starting again with the original urn. It is easy to see that exactly the same factors that appear in the above expression will appear again, except that the factors in the numerator will appear in a different order. Hence each possible sequence of r_1 of the b_1's and r_2 of the b_2's has the same probability, which can be expressed as

$$\left\{\prod_{i=1}^{2}\Gamma(n_i+r_i)/\Gamma(n_i)\right\}/\{\Gamma(n+R)/\Gamma(n)\}$$

where $n = n_1 + n_2$. This fundamental property of the Polya urn model makes it easy to use. It follows that the marginal probability of getting a ball coloured b_1 on any draw is just n_1/n. Moreover, the joint probability distribution for any pair of draws is the same as that for any other pair of draws, and the same is true for triples and so on. In fact the distribution of the draws is invariant under any permutation of the draws and hence the underlying distribution is exchangeable. After R draws, when R is large, it is natural to wonder about the composition of the urn. Let W_R be the proportion of balls with colour b_1 in the urn after R draws. Then W_R is a random variable which takes on values in the unit interval and its limiting distribution, as R approaches infinity, is beta(n_1, n_2).

These facts generalize in a straight forward way to urns con-

taining balls of k different types. Suppose for $i = 1, \ldots, k$ the urn contains n_i balls each labelled b_i where the b_i's are k distinct real numbers. Let $n = \sum_{i=1}^{k} n_i$. Then consider the experiment where a ball is chosen at random from the urn and then it and one more of the same type are returned to the urn. Suppose this experiment is repeated R times. Then the resulting joint probability distribution is exchangeable and the probability of getting r_i balls of type b_i in some specified order is

$$\left\{ \prod_{i=1}^{k} \Gamma(n_i + r_i)/\Gamma(n_i) \right\} / \{\Gamma(n + R)/\Gamma(n)\} \tag{2.12}$$

where $R = \sum_{i=1}^{k} r_i$. Furthermore the limiting distribution of the vector of proportions of each type of ball in the urn after R draws is Dirichlet $(n_1, \ldots, n_k)$, as R goes to infinity. This limit result is related to de Finetti's theorem. For an elementary proof of this theorem see Heath and Sudderth (1976). A good discussion of the properties of the Dirichlet distribution can be found in Wilks (1962).

2.6 The Polya posterior

In Section 2.3.2 we proved the admissibility of the sample mean for estimating the population mean, under squared error loss, by showing that it was a stepwise Bayes estimator against a sequence of priors. Given any possible data point there was a 'posterior distribution' for which the sample mean was the Bayes estimate for that particular point. This did not show that the sample mean is a Bayes estimator, however, because there is no single prior distribution against which the pseudo 'posterior distributions' are the actual posterior distributions for all of the sample points. For further discussion on this point see Muliere and Secchi (1996). In this section we will show that these 'posterior distributions' are in fact Polya urn distributions and discuss some of the implications of this for inference in finite population sampling.

For definiteness we suppose our data point $z = (s, z_s)$ is such that $n(s) = n$ and s just consists of the first n units of the population. Let $\mathbf{b} = (b_1, \ldots, b_k)^T$ be the k distinct values that appear in $(z_1, \ldots, z_n)^T$ and for $i = 1, \ldots, k$ let $n_i = c_z(i, s)$ be the number of z_j's in z_s which equal b_i. Therefore $n_i \geq 1$ and $\sum_{i=1}^{k} n_i = n$. Let $\mathbf{b}^o$ be any vector of distinct real numbers, of length at least k,

whose values contain all the values of $\mathbf{b}$. In the proof of Theorem 2.3, when the parameter space is $\mathcal{Y}(\mathbf{b}^o)$, the data z is considered in the stepwise Bayes argument at the step when the subset of the parameter space is $\mathcal{Y}^*(b_1, \ldots, b_k)$. On this set we have by equation (2.7) that the prior is

$$\pi(\mathbf{y}) \propto \int_0^1 \cdots \int_0^1 \prod_{i=1}^{k-1} \theta_i^{c_y(i)-1} \left(1 - \sum_{i=1}^{k-1} \theta_i\right)^{c_y(k)-1} d\theta_1 \cdots d\theta_{k-1}. \tag{2.13}$$

We now wish to compute the conditional probability, given our sample point z, that for the unseen $\{ y_j : j = n+1, \ldots, N \}$ there are exactly r_i of the b_i's, for $i = 1, \ldots, k$ in some specified order, where $\sum_{i=1}^{k} r_i = N - n$. Let $\mathbf{y}'$ be such a point which is also consistent with the observed data, i.e. $\mathbf{y}'(s) = z_s$. Then we wish to find

$$\begin{aligned} \pi(\mathbf{y}'|z) &= \frac{\int_0^1 \cdots \int_0^1 \prod_{i=1}^{k-1} \theta_i^{n_i+r_i-1} (1 - \sum_{i=1}^{k-1} \theta_i)^{n_k+r_k-1} \, d\theta_1 \cdots d\theta_{k-1}}{\int_0^1 \cdots \int_0^1 \prod_{i=1}^{k-1} \theta_i^{n_i-1} (1 - \sum_{i=1}^{k-1} \theta_i)^{n_k-1} \, d\theta_1 \cdots d\theta_{k-1}} \\ &= \left\{ \prod_{i=1}^{k} (\Gamma(n_i + r_i)/\Gamma(n_i) \right\} / \{\Gamma(N)/\Gamma(n)\}. \end{aligned} \tag{2.14}$$

Note that this is exactly the same formula that is given in (2.12) for the probability of seeing exactly r_i of the b_i's, for $i = 1, \ldots, k$ in some specified order for $N - n$ future draws from an urn, which for each i contains n_i of the b_i's. Hence given an observed data point z the pseudo 'posterior distribution', arising from the stepwise Bayes argument, for the unseen given the seen is just what we would get if we placed the observed units in an urn and did Polya sampling from the urn to assign values for all of the unseen units. We shall now argue that this pseudo 'posterior distribution' is just not a trick to prove some admissibility results but yields a noninformative Bayesian approach to finite population sampling. We will call this pseudo 'posterior distribution' the **Polya posterior**.

We believe that the most sensible approach to finite population sampling is a conditional one. After the data have been observed our inferences should just depend on the single realized sample values; that is, as we argued in Chapter 1, we believe in the likelihood

principle as applied to finite population sampling. (Some frequentists use the term 'conditional' when the conditioning may be on a subset of the possible samples, of which the realized sample is just one instance. This is not what we mean by a conditional approach to finite population sampling.) We shall argue that a conditional approach to finite population sampling is possible even though one cannot fully specify a prior distribution. In particular, a conditional approach is possible even when little prior information is present.

Suppose, for example, that we wish to estimate the population mean in the presence of scant prior information. We choose as our design simple random sampling. For this design the sample mean is an unbiased estimator of the population mean. In this setting some who find the appeal to unbiasedness unconvincing still believe the sample mean to be a sensible estimate of the population mean. Hence it is of interest to find a conditional justification for the sample mean for this setup.

As we remarked before, the basic problem of finite population sampling is deciding what one learns about the unobserved units from the observed sampled units. The usual justification of the sample mean, based on unbiasedness, is not very appealing to one who prefers to think conditionally. Naively, almost everyone thinks conditionally to some extent, we believe, since almost no one would be comfortable using the sample mean as an estimate if they had some reason to believe that their observed sample was not 'representative' of the population as a whole. Although the notion of a representative sample has an intuitive appeal, it is seldom given a precise definition. For our purposes we say that a sample is representative if in our judgement the 'seen' and 'unseen' units are roughly exchangeable. The next step is to find a 'posterior distribution' which reflects such a judgement. An appropriate distribution must treat the unseen units exchangeably and assign each of them a marginal distribution that closely mimics the seen. Note this is exactly what the Polya posterior does.

To see why the Polya posterior is an appropriate model for our beliefs, after we have observed a sample, where in our judgement the seen and the unseen are roughly exchangeable, consider the following. Suppose that we are choosing at random, without replacement, successive members of a population about which we have little prior information. After we have observed n units what should our beliefs be about the next item to be drawn? If we believe the sample is representative, then a reasonable distribution for the

next item chosen is just the empirical distribution of the observed sample up to that point. By the same argument, after the next item is observed the empirical distribution based on the first $n+1$ units will be a reasonable guess for the distribution of the $(n+2)$th item to be chosen. But this inductive argument suggests the Polya posterior. When the sample size is small, clearly almost no one would find the Polya posterior an acceptable description of their beliefs. For larger sample sizes, however, it seems to capture the idea that the unseen are like the seen and gives an intuitive justification for the sample mean as an estimate of the population mean. Note that the design probabilities play no role in the above argument and all that is needed is for the statistician to be able to assert the approximate exchangeability of the seen and the unseen. Intuitive justifications have their place but the more important question is 'Does the Polya posterior lead to sensible methods in practice?' We already know that the Polya posterior will give admissible point estimators for any parameter of interest. In the rest of this chapter we will show that it also yields good interval estimators and estimates of variability. In summary, we will argue that the Polya posterior gives a noninformative Bayesian or conditional approach to finite population sampling which yields procedures with good frequentist properties for a variety of problems.

To help see why this is so, consider again the problem of estimating the population mean, $\mu = \mu(\mathbf{y})$. Recall, that given an observed data point $z = (s, z_s)$, we have under the Polya posterior that $E(\mu|z) = \bar{z}_s$, the sample mean since for $j \notin s$, $E(y_j|z) = \bar{z}_s$. We now want to calculate $\mathrm{Var}(\mu|z)$, the posterior variance of μ given z under the Polya posterior. Let $\mathrm{Var}(z) = (n(s)-1)^{-1}\sum_{i\in s}(z_i - \bar{z}_s)^2$ be the sample variance. In the following calculation we will assume for notational simplicity that the units 1 and 2 do not belong to the sample s and $n = n(s)$. Then, remembering that under the Polya posterior the marginal distribution of any unseen y_j assigns probability $1/n$ to each unit in the sample, we have that

$$N^2\mathrm{Var}(\mu|z) = \mathrm{Var}\left(\sum_{j\notin s} y_j|z\right)$$
$$= \sum_{j\notin s}\mathrm{Var}(y_j|z)$$

$$
\begin{aligned}
&+ \sum_{i<j\,:\,i\notin s \text{ and } j\notin s} 2\,\mathrm{Cov}((y_i,y_j)|z) \\
&= (N-n)\mathrm{Var}(y_1|z) + 2\binom{N-n}{2}\mathrm{Cov}((y_1,y_2)|z) \\
&= (N-n)(E(y_1^2|z) - \bar{z}_s^2) \\
&+ (N-n)(N-n-1)(E(y_1y_2|z) - \bar{z}_s^2) \\
&= (N-n)\left(\frac{\sum_{i\in s} z_i^2}{n} - \bar{z}_s^2\right) \\
&+ (N-n)(N-n-1)(E[E(y_1y_2|y_1,z)|z] - \bar{z}_s^2) \\
&= (N-n)\frac{n-1}{n}\,\mathrm{Var}(z) \\
&+ (N-n)(N-n-1)\left(E[y_1\frac{y_1+\sum_{i\in s} z_i}{n+1}|z] - \bar{z}_s^2\right) \\
&= (N-n)\frac{n-1}{n}\,\mathrm{Var}(z) \\
&+ (N-n)(N-n-1)\left(\frac{\sum_{i\in s} z_i^2}{(n+1)n} + \frac{\bar{z}_s\sum_{i\in s} z_i}{n+1} - \bar{z}_s^2\right) \\
&= N(N-n)\frac{\mathrm{Var}(z)}{n}\frac{n-1}{n+1}
\end{aligned}
$$

where the last step follows from the definition of $\mathrm{Var}(z)$ and some simple algebra. This calculation is well known, see for example page 46 of Rubin (1987) and Ericson (1969b). If we let $f = n/N$ then we can rewrite the above as

$$\{\mathrm{Var}(\mu|z)\}^{\frac{1}{2}} = \frac{\mathrm{Var}(z)^{\frac{1}{2}}}{\sqrt{n}}(1-f)^{\frac{1}{2}}\left\{\frac{n-1}{n+1}\right\}^{\frac{1}{2}}. \tag{2.15}$$

Furthermore, it is known (Lo, 1988) that for any z with a large n that given z the distribution of μ under the Polya posterior is approximately normal with mean $\bar{z}_s$ and variance given above. But the usual classical confidence interval for μ, see page 27 of Cochran (1977), is based on the assumption that the sample mean, $\bar{z}_s$, is approximately normally distributed with mean μ and estimated standard deviation $(\mathrm{Var}(z)(1-f)/n)^{1/2}$. This suggests that for reasonably large sample sizes, intervals for the mean based on the Polya posterior should closely approximate the usual frequentist intervals. Note that the underlying logic justifying these two very similar but different intervals is quite dissimilar since the design probabilities, which are crucial for the frequentist argument, played

no role in the development of the Polya posterior. In the rest of this chapter we will consider the behaviour of interval estimators, based on the Polya posterior in a variety of problems and compare them to standard frequentist intervals.

2.7 Simulating the Polya posterior

In this section we will show how inference procedures based on the Polya posterior can be found approximately by simulation. We have seen that the Polya posterior estimator of the population mean, under squared error loss, is just the sample mean. But this is the only estimator of interest, based on the Polya posterior which can be found in closed form. When estimating the median or when interval estimates of the median or mean are desired, one needs to simulate the Polya posterior to find the estimators approximately. In Section 2.6 we saw how the pseudo posterior arising from the stepwise Bayes argument for admissibility could be identified with a Polya urn model. This not only led to the name, Polya posterior, but gives an easy way to simulate this distribution. Given a sample point $z = (s, z_s)$ one just puts the values of the $n(s)$ observed units in an urn and then makes $N - n(s)$ Polya draws from the urn. Hence, when we stop, the urn contains a total of N units. This yields one simulated copy of the entire population. If for the given z this is repeated R times we then have produced R simulated copies from the Polya posterior of the whole population. If R is large then we can find approximately almost any characteristic of the posterior that we desired.

For example suppose that $\gamma(\mathbf{y})$ is a function of the parameter for which we want a point estimate based on the Polya posterior. For a given z we will simulate the Polya posterior R times. Usually R equal to 500 or 1000 is sufficient. For each simulated copy of the population we will compute the value of $\gamma(\cdot)$. If the loss function is squared error we would then compute the mean of the R simulated values of $\gamma(\mathbf{y})$ and this would give the approximate value of the stepwise Bayes estimate with respect to the Polya posterior. If the loss function is absolute error we would find the median of the R simulated values of $\gamma(\mathbf{y})$. So we see that if we can simulate the Polya posterior, the corresponding point estimate for many different functions $\gamma(\mathbf{y})$ can be found approximately for a variety of loss functions.

Unfortunately the problem of finding set estimates is more com-

plicated. Suppose a Bayesian wishes to announce a credible set with posterior probability 0.95 for a real valued continuous parameter. Standard practice is to choose the highest posterior density region, see Berger (1985). If the posterior is unimodal such a region must necessarily be an interval. Intervals are preferred since they are easier to interpet than more complicated regions. Following frequentist practice a Bayesian could choose some fixed number between 0 and 1, say 0.95, and then given the data announce the 0.95 highest posterior density region. In Meeden (1990) it was shown that in some cases the resulting set estimation procedure is inadmissible. That is, fixing the amount of posterior probability and announcing the smallest set with this much probability can lead to inadmissible procedures.

An alternative method would be to fix some positive constant and announce the set consisting of all those parameter points where the posterior density function exceeds the selected constant. Note that the posterior probability of this set can vary with the data. Any Bayesian set estimation procedure can be thought of as a crude summary of the posterior distribution. This second approach is analogous to a contour curve of a map. So even though the second is widely used in other areas, the first would be the preferred approach for most Bayesian statisticians.

For discrete problems with a finite parameter space we showed in Section 2.4 that the second approach gives admissible set estimation procedures in the stepwise Bayes setting. Hence, if in finite population sampling we assume the parameter space to be $\mathcal{Y}(\mathbf{b})$ for some vector $\mathbf{b}$, Theorem 2.5 guarantees that the second approach will yield admissible set estimators when used in conjunction with the Polya posterior. Unfortunately each announced set will just be a finite collection of points. Not only will such a set be difficult to interpret but also difficult to find approximately. Moreover it will be difficult to compare to standard frequentist methods, since the amount of posterior probability assigned to the announced set can vary with the observed data. For these reasons we will use a more naive set estimate based on the Polya posterior. In some cases it can be thought of as a reasonable approximation to the admissible procedure described above.

Suppose that $\gamma(\mathbf{y})$ is the function for which we want an interval estimate. For a given sample we will simulate the Polya posterior distribution of $\gamma(\mathbf{y})$ by creating R simulated copies of the entire population and computing the value of $\gamma(\cdot)$ in each case. For this

simulated population of values we let $q025$ and $q975$ be the 0.025 quantile and 0.975 quantile, respectively. Then $(q025, q975)$ will be our announced set estimate, and it will always have approximate 'posterior probability' 0.95 under the Polya posterior. This interval is easy to compute and in keeping with standard Bayesian practice. In the following this interval will be compared to the usual frequentist 95% confidence interval for $\gamma(\mathbf{y})$. For convenience we will limit ourselves to sets with posterior coverage probability of 0.95, since this seems to be the popular value of a nominal confidence level in practice.

It remains to describe how we generate 'Polya samples from an urn'. For definiteness suppose that we have an urn containing n items. Attached to each item is the value of a characteristic y, possibly vector valued. Let k be the number of distinct y values occurring, and let n_i be the number of items in the urn that have the ith value. Let $b_1, \ldots, b_k$ denote these k distinct values of the characteristic y appearing in the urn. Clearly, $n_i \geq 1$ for $i = 1, \ldots, k$, and $\sum_{i=1}^{k} n_i = n$. Suppose now that we wish to generate a Polya sample of size $M = N - n$, the population size minus the sample size.

Here is one method that we have found useful. Label the n items in the urn from 1 to n in some specified manner; the particular order chosen is not important. Observe a uniform random variable over the set $\{1, 2, \ldots, n\}$. For the observed i find the y value that is labelled i, create a new item with this value, and label it $n+1$. Next observe a uniform random variable over the set $\{1, 2, \ldots, n, n+1\}$ to get as before a new item labelled $n+2$. Continue in this way until our Polya sample of size M is completed. Let $\tilde{n}_i$ denote the total number of items in the urn with the value b_i, after the Polya sample is completed. Note $\tilde{n}_i \geq n_i$, for $i = 1, \ldots, k$, and $\sum_{i=1}^{k} \tilde{n}_i = N$. Let $\tilde{p}_i = \tilde{n}_i / N$. Hence we have sampled one particular realization of the entire population under the Polya posterior. For many parameters of interest, $\gamma(\mathbf{y})$, one can compute the value for the realization just knowing the b_i's and the $\tilde{p}_i$'s. One repeats this process R times to approximate the distribution of $\gamma(\mathbf{y})$ under the Polya posterior. One then uses this simulated distribution to find approximately the estimate of interest.

We do not claim that this is necessarily the fastest or most efficient way to do Polya simulations. We only note that this method is easy to use even for quite large values of R, say in the thousands. For large R there is an alternative method for obtaining an ap-

proximation based on asymptotic theory. It is well known that for large R the distribution of the $\tilde{p}_i$'s is approximately Dirichlet with parameters $n_1, \ldots, n_k$. Hence rather than generate a Polya sample one can sample from the appropriate Dirichlet distribution to get a realization of the population relative frequencies under the Polya distribution. But a realization of a Dirichlet random vector can be found by observing k independent gamma random variables, say $W_1, \ldots, W_k$, where W_i has shape parameter n_i and they all have scale parameter 1, and letting $\tilde{p}_i = W_i / \sum_{j=1}^{k} W_j$.

In the results that follow we have used both of these methods. For either method it remains to choose R, the number of Polya copies of the entire population used in the approximation. To some extent, an appropriate choice depends on k and the n_i's. In most of the examples we chose $R = 500$. In some cases we compared the results for $R = 500$ and $R = 1000$ and observed that they were essentially the same. Because of the large number of simulations we ran, we preferred to use the smaller value whenever possible. In practice for a particular problem one can compare the results for increasing values of R and stop when the resulting approximations appear to converge.

2.8 Some examples

In this section we will compare point and interval estimators based on the Polya posterior to some frequentist estimators for several different problems. In most of the cases the estimators based on the Polya posterior will be found approximately using the methods discussed in the previous section.

2.8.1 Set estimation for the mean and median

Let the approximate Polya posterior interval discussed in the previous section be denoted by AP. When estimating the mean it will be compared to the usual normal theory interval (NT) of Cochran (1977). When estimating the median an interval due to Woodruff (1952) has often been used. Let this interval be denoted by W. Recently, Francisco and Fuller (1991) considered an alternate set estimator of the median. In the simplest situation it is easy to describe. Let F denote the distribution function which puts mass $1/N$ at each y_i in the population. For a fixed v, the proportion of members of a sample of size n, less than or equal to v is approxi-

mately binomial $(n, F(v))$. One can then find an approximate 95% confidence interval for the median as follows. For a given sample one selects all those v's for which the usual interval estimate for $F(v)$, which is based on the normal approximation to the binomial distribution, contains 1/2. Francisco and Fuller showed that this test inversion procedure, which we denote by FF, compares favourably to W. So when estimating the median we will compare AP to both FF and W.

The first population to be considered is one that was discussed in Greenlees *et al.* (1982). It consists of wages and salaries of 5515 individuals. As is typical for such populations of income data, it is skewed to the right. For samples of sizes 20, 40 and 60, approximate 0.95 probability intervals based on the Polya posterior were computed. This was done for both the mean and the median and are denoted by AP. They were compared to the usual 95% normal theory confidence intervals for the mean, NT, and to the 95% FF intervals for the median. The results are summarized in Table 2.1. We see that the AP behaviour is similar to that of NT (as was to be expected). The AP intervals are just slightly shorter on the average than the NT intervals and cover the true value of the mean a bit less often. However, their behaviour becomes more similar as the sample size increases. For estimating the median there is little difference between the performances of the AP intervals and the FF intervals.

We did comparisons for ten other populations. Each population contained 500 units. ppexp was a random sample from an exponential distribution with parameter equal to 1. ppc was a random sample from a Cauchy distribution centred at 0 with scale parameter 1. ppln was a random sample from a lognormal distribution with mean and standard deviation (of the log) 4.9 and 0.586 respectively. ppn1 contained a random sample of size 400 from a standard normal distribution and a random sample of size 100 from a uniform (−0.01, 0.01) distribution. ppn2 was a random sample of 450 from a standard normal distribution and 50 units equal to 2.5. ppu1 was a random sample of size 400 from a uniform (−1, 1) distribution and a random sample of size 100 from a uniform (−0.01, 0.01) distribution. ppu2 consisted of 300 uniform (−1, 1) observations, 100 uniform observations (−0.5, 0.5) observations and 100 uniform (−0.1, 0.1) observations. ppu3 consisted of 200 uniform (−1, −0.5) observations, 100 uniform (−0.5,0.5) observations and 200 uniform (0.5, 1) observations. ppu4 consisted of 100 uniform (−1,

Table 2.1 *Comparison of 95% frequentist intervals and 0.95 credible intervals based on the Polya posterior for the mean and median for the population of income data.*

	Average length	Relative frequency of coverage
	3000 samples of size 20	
Mean		
NT	65.39	0.898
AP	62.42	0.885
Median		
FF	59.63	0.963
AP	57.56	0.957
	1500 samples of size 40	
Mean		
NT	48.83	0.937
AP	46.04	0.930
Median		
FF	40.23	0.959
AP	37.30	0.950
	1750 samples of size 60	
Mean		
NT	38.78	0.919
AP	38.48	0.919
Median		
FF	29.72	0.943
AP	29.67	0.949

1) observations, 200 uniform $(-1, -0.99)$ observations and 200 uniform (0.99, 1) observations. Finally the last population comes from sunspot data discussed in Andrews and Herzberg (1985). These data are successive monthly means of daily sunspot numbers beginning in the year 1749. Let ppssp5 denote the first 500 numbers of these data. This population is skewed strongly to the right with mean = 59.60, median = 50.55 and variance = 1808.53.

For each population we considered estimation of the median based on samples of size 50. As a first step we estimated the variance of the sample median using 500 simple random samples of

size 50. From this we calculated the 'true average length' of a 95% confidence interval for the population median as 2(1.96) times the square root of this variance. For each population this true average length is given in column 1 of Table 2.2. We then took 500 simple random samples (without replacement) of size 50 and for each sample calculated the AP interval based on 500 Polya simulations of the entire population. For these 500 intervals, an average length was found and the frequency of coverage was computed. These results are given in the second column of Table 2.2. For a second group of 500 samples of size 50 we calculated the average length and frequency of coverage for the FF intervals, the W intervals and for an adjusted version of the FF intervals, denoted here by FFc. These results are given in the last three columns of Table 2.2. We also compared the AP and NT intervals for the mean. In all cases the two sets of intervals for the population mean were quite similar and hence we have not included any of those results.

Note that in every case the AP interval is longer on average than the FF interval and the FF interval seems to be 'undercovering'. Now it is easy to check that for samples of size 50 the FF interval is just $(z_{(19)}, z_{(31)})$ where $z_{(i)}$ is the value of the ith order statistic of the sample when all of the values in the sample are different. Since the upper limit of the approximate 95% confidence interval, based on the binomial distribution for $F(v)$ when $v = z_{(19)}$ is 0.5145 and for $F(v)$ when $v = z_{(18)}$ is 0.493, we should be able to improve the performance of the FF interval by a continuity correction. This leads to the interval with lower limit $0.67z_{(18)} + 0.33z_{(19)}$ and with a similar adjustment on the upper limit. This is the interval denoted by FFc, and in most cases it is preferred to FF. Overall, the behaviours of the AP intervals and the FFc intervals seem quite comparable. The W intervals are given by $(0.93z_{(18)} + 0.07z_{(19)},\ 0.07z_{(31)} + 0.93z_{(32)})$, when a similar continuity correction is used. As we see from the last column of Table 2.2, the W intervals seem to be a bit too long and tend to 'overcover'. Finally, we note that the difference between FF and FFc depends on the sample size n. When $n = 60$, for example, the lower limit of the FF is $z_{(23)}$, whereas for FFc it is $0.35z_{(22)} + 0.65z_{(23)}$. This smaller adjustment indicates why the 'undercoverage' problem of FF did not appear in Table 2.1.

Table 2.2 *The average length and relative frequency of coverage for some 95% frequentist intervals and the 0.95 credible interval based on the Polya posterior for the median for a sample size of 50.*

Population	'Length'	AP	FF	FFc	W
ppexp	0.567	0.538	0.490	0.542	0.562
		0.950	0.948	0.974	0.978
ppc	0.793	0.905	0.786	0.882	0.920
		0.942	0.924	0.960	0.962
ppln	58.638	57.673	52.964	59.210	61.633
		0.944	0.926	0.948	0.950
ppn1	0.182	0.313	0.260	0.320	0.344
		0.946	0.926	0.952	0.966
ppn2	0.667	0.686	0.630	0.703	0.732
		0.950	0.932	0.960	0.964
ppu1	0.146	0.209	0.184	0.228	0.245
		0.958	0.934	0.956	0.964
ppu2	0.198	0.214	0.200	0.228	0.240
		0.950	0.914	0.940	0.950
ppu3	1.237	0.960	0.922	0.988	1.014
		0.940	0.906	0.948	0.954
ppu4	2.397	1.715	1.637	1.718	1.749
		0.938	0.920	0.948	0.958
ppssp5	35.171	30.397	28.001	31.112	32.319
		0.930	0.920	0.940	0.946

2.8.2 Point estimation of the median

As we have seen, under squared error loss the Polya posterior estimator of the population mean is just the sample mean. Suppose now we wish to estimate the population median. Given a sample, the Polya posterior induces a predictive distribution for the population median. Two natural estimators for the population median are the mean and median of this predictive distribution. Neither of these are easy to compute directly. As we noted in Section 2.5, it is well known that, when $N - n$ is large the Polya posterior distribution of the population given the sample z is approximately Dirichlet. Following an argument on page 224 of Ferguson (1973) we have in this case that the median of the predictive distribution

for the population median is approximately the sample median. Hence in what follows we will compare the mean of the predictive distribution for the population median, denote by **mp-median**, and the sample median, denoted by **s-median**.

Since the mean is the Bayes estimate under squared error loss and the median is the Bayes estimate under absolute error loss, one might expect that the estimator mp-median would perform better than the estimator s-median for squared error loss and vice versa for absolute error loss. To see if this was indeed the case, we compared these two estimators for the ten populations studied in Table 2.2. For each population we took 500 simple random samples of size 25 and 500 more of size 50. For each sample we computed both estimators. (We found mp-median approximately, based on 500 Polya simulations of the entire population.) Then their losses under squared error and absolute error were found. In Table 2.3 we give the quotients of the average error of s-median divided by the average error of mp-median; that is, mp-median is preferred to s-median whenever the stated ratio is greater than 1.

We see that for seven of the populations mp-median is the preferred estimator for both loss functions and s-median is the preferred estimator for the other three populations, again for both loss functions. Note that these three ppu1, ppu2 and ppn1, are quite artificial in nature, since 20% of the population lies very close to the population median, which essentially guarantees that s-median will be a good estimate for the population median. Because of the robust nature of s-median it essentially ignores much of the information contained in the sample. Hence if we are in a situation where most of the population lies some distance from the population median, as in ppu3 and ppu4, we would expect to be able to improve on the estimator s-median by incorporating into our estimator more of the information contained in the sample. This is, in fact, what the estimator mp-median does, since it depends more fully on all the actual values in the sample than does s-median. What is perhaps more surprising is how much better mp-median performs than s-median for the more typical populations. We also see from Table 2.3 that the amount of improvement of mp-median over s-median seems to decrease as the sample size increases. This seems to be consistent with the preceding explanation of why such an improvement should exist. These results suggest that mp-median is preferred to s-median as a point estimator of the population median, except when the population is

Table 2.3 *The ratio of the average squared error and average absolute error for the estimators s-median and mp-median for the population median.*

Population	Sample size	Ratio of average squared errors	Ratio of average absolute errors
ppexp	25	1.31	1.17
	50	1.15	1.08
ppc	25	1.03	1.00
	50	1.10	1.02
ppln	25	1.21	1.16
	50	1.17	1.07
ppn1	25	0.89	0.63
	50	0.69	0.49
ppn2	25	1.14	1.05
	50	1.09	1.04
ppu1	25	0.76	0.55
	50	0.53	0.45
ppu2	25	0.88	0.90
	50	0.93	1.00
ppu3	25	1.69	1.31
	50	1.36	1.13
ppu4	25	1.92	1.45
	50	1.63	1.31
ppssp5	25	1.17	1.09
	50	1.15	1.07

highly concentrated about its median or the sample size is quite large. In all these examples both estimators were approximately unbiased, although the mean of s-median is usually a bit closer on the average to the population median than that of mp-median.

2.8.3 Estimating the ratio of two medians

One of the advantages of the Polya posterior is that it is easy to implement in both simple and more complex situations. Consider the problem of estimating the ratio of means or the ratio of medians for two related populations (i.e., a bivariate population). For example, the two populations might just be the same population

at two different time periods in a situation where one observes the same set of units in each sample. For such a sample the Polya posterior assumes that pairs of y values for the unobserved units are exchangeable given the pairs of y values for the observed units. Hence given such a sample the Polya posterior is just like Polya sampling from an urn with n items where n is the number of units that appear in the sample. Moreover, each item in the urn would have two values associated with it, the y values from the two different time periods. Polya sampling would then create related but separate copies of the two populations and the AP intervals for the ratio of the means or the ratio of the medians can easily be found. For the ratio of means, an approximate confidence interval is well known; see, for example, page 129 of Jessen (1978). For the ratio of the medians there seems to be no readily available frequentist interval. In what follows we will compute the AP intervals for the ratio of medians for seven different populations.

Our bivariate populations, denoted by (y_i', y_i) for $i = 1, \ldots, N$, will all be constructed in a similar manner. First we will specify a univariate population for the y_i' values. Then for each i we will specify the conditional distribution of y_i given y_i'. For $i \neq j$ these conditional distributions will be independent. Finally, we will generate an actual set of y_i's according to these distributions.

The population, ppstsk, consists of 1000 units and is strongly skewed to the right. Its mean is 42.63, its median is 39.29 and its variance is 204.59. From this univariate population of y_i''s values we will construct three bivariate populations. In the first ppstska, y_i was set equal to an observation from a normal distribution with mean $y_i' + 5$ and variance y'. For the resulting bivariate population the correlation is 0.91 and the ratio of the median of y to the median of y' is 1.14. In ppstskb, y_i was set equal to an observation from a normal population with mean $y_i' + 5$ as before, but now with variance $9y_i'$. In this case the resulting population has correlation 0.61 and the ratio of the medians is 1.17. In population ppstskc, y_i was set equal to an observation from a normal distribution with mean $y_i' + 5$ and variance $(y_i')^2$. The correlation between y and y' is 0.33 and the ratio of the medians is 1.09.

The population ppsk, also consists of 1000 units. It is less skewed than ppstsk and has heavier tails. Its mean is 102.19, its median is 100.80 and its variance is 459.26. From this population we created a bivariate population ppska by setting y_i equal to an observation from a normal random variable with mean $6\sqrt{y_i'}$ and variance

$(0.45)^2 y_i'$. The correlation between y and y' is 0.82 and the ratio of the median of y to the median of y' is 0.60.

The population ppln is just the sample from a lognormal population described in Section 2.8.1. To create the bivariate population pplna, y_i was set equal to an observation from a normal distribution with mean $y_i' + 2 \log y_i'$ and variance $(y_i')^2$. This resulted in 47 units with negative y values. Since it is usually the case in such problems for each variable to be strictly positive we replaced these 47 negative values with a random sample from the standard exponential distribution. For this modified population, the correlation between y and y' is 0.58 and the ratio of the median of y to the median of y' is 1.27.

The population ppexp1 is just the population ppexp, described in Section 2.8.1 with 5 added to each unit. To create the bivariate population ppexp1a, y_i was set equal to an observation from a normal distribution with mean $y_i' + 5$ and variance $4(y_i' - 5)$. The correlation between y and y' is 0.49 and the ratio of the median of y to the median of y' is 1.85.

The population ppssp5 was also described in Section 2.8.1. To create the bivariate population ppssp5a, y_i was set equal to an observation from a normal distribution with mean $500 - 2y_i'$ and variance $(y_i')^2$. The correlation between y and y' is -0.77 and the ratio of the median of y to the median of y' is 8.06.

For each of the seven bivariate populations, an estimate of the variance of the ratio of the sample medians was found based on 500 samples. This gives the 'true length' of a NT 95% confidence interval for the ratio of the medians found in Table 2.4. The rest of the table summarizes the performance of the AP intervals. Note that in some cases the AP intervals tend to 'overcover'. In all but one case the average length of the AP intervals is less than 10% longer than the 'true length', and in the worst case exceeds it by less than 15%. Hence for the problem of interval estimation of the ratio of two medians the Polya posterior seems to give sensible answers, even though in some instances they tend to be a bit conservative.

We did not include any results for the ratio of the means, since the performances of the AP intervals and classical intervals are quite similar. For example, for 500 random samples of size 75, when estimating the ratio of means in ppstska the approximate classical interval's relative frequency of coverage was 0.954 and its average length was 0.0714. The AP interval's relative frequency of coverage was 0.952 and its average length was 0.0709.

Table 2.4 *Performance of 0.95 AP intervals for the ratio of two medians.*

Population	Sample size	'True length'	Average length	Relative frequency of coverage
ppstska	25	0.282	0.321	0.988
	75	0.154	0.176	0.994
ppstskb	25	0.506	0.579	0.958
	75	0.289	0.305	0.954
ppstskc	25	0.876	0.915	0.966
	50	0.640	0.635	0.956
ppska	25	0.085	0.096	0.982
	50	0.059	0.065	0.982
pplna	25	1.130	1.144	0.986
	50	0.736	0.800	0.972
ppexp1a	25	0.304	0.334	0.976
	50	0.226	0.227	0.980
ppssp5a	25	9.541	10.263	0.942
	50	6.320	6.388	0.946

Next we consider point estimators of the ratio of two medians. One natural estimator is just the ratio of the sample medians. On the other hand, given a sample one can produce simulated Polya realizations of the whole bivariate population. For each such realization one can find the ratio of the medians and then take as an estimate the mean of these simulated values of the ratio.

In Table 2.5 we compare the average squared error losses and the average absolute error losses of this Polya posterior point estimator to those of the natural estimator, the ratio of the two sample medians. In all but the last case the Polya posterior point estimator is clearly preferred, since all the ratios are greater than 1. In all of the cases, except the last one, both of the estimators are approximately unbiased. Recall that for population ppssp5a the ratio of the two medians is 8.06. For sample size 25, 8.81 and 8.46 are the average values of the two estimators, whereas for sample size 50, 8.41 and 8.29 are the two average values. In each case the Polya posterior point estimator is more biased.

In Tables 2.2, 2.3, 2.4 and 2.5, for a given sample, the Polya posterior procedures were found using 500 Polya simulations. In two

Table 2.5 *The ratio of the average squared error and average absolute error losses for the quotient of the sample medians and the Polya posterior estimator when estimating the quotient of two medians.*

Population	Sample size	Ratio of average squared errors	Ratio of average absolute errors
ppstska	25	1.94	1.34
ppstskb	25	1.33	1.14
ppstskc	25	1.28	1.10
ppska	25	1.48	1.21
pplna	50	1.20	1.08
	25	1.43	1.22
ppexpla	25	1.30	1.15
ppssp5a	50	1.01	1.03
	25	0.95	1.02

examples, one from Table 2.3 and one from Table 2.4, we repeated the simulations using 1000 Polya simulations for each sample. In each case the results were essentially indistinguishable from the earlier simulations based on 500 Polya realizations.

CHAPTER 3

Extensions of the Polya posterior

In Chapter 2 we argued that the Polya posterior was a sensible non-informative Bayesian appproach to problems in finite population sampling when one's beliefs about the units are exchangeable. The underlying theoretical justification was a stepwise Bayes argument that demonstrated the admissibilty of the resulting procedures. In this chapter we will consider some extensions of the Polya posterior which allow one to incorporate different kinds of prior information into the analysis. Since the theorectical justification for these methods is again a stepwise Bayes one, a full-blown prior need not be specifed. However it will be the case that for these problems the inferences will be done in the usual Bayesian manner and will be based on a 'pseudo posterior'.

In Section 3.1 we consider situations where either a prior guess for the entire population or a prior guess for each unit is available. In Section 3.2 we consider the situation where an auxiliary variable is available and one's beliefs about the ratios are exchangeable. In particular we consider the problem of finding an interval estimate of the population median. In Section 3.3 we study Bayes and pseudo Bayes estimators in stratified populations with two different levels of prior knowledge about the stratification. In Section 3.4 we consider the problem where we must not only select an estimator but also must select a design from some class of possible designs. We show that the stepwise Bayes technique can be used to identify uniformly admissible pairs of estimators and designs. In Section 3.5 the problem of nonresponse will be considered. We will show how the Polya posterior along with an assumed model for the relationship between the responders and the nonresponders leads to procedures similar in spirit to those arising from multiple imputation (Rubin, 1987). In Section 3.6 we will see that there is a close relationship between admissibility questions in finite population sampling and admissibility questions in nonparametric problems. We will show that the stepwise Bayes argument which gives the

admissibility of the procedures based on the Polya posterior can easily be adapted to yield the admissibility of standard nonparametric estimators. The relationship between the Polya posterior and the Dirichlet process priors (Ferguson, 1973) and the Bayesian bootstrap (Rubin, 1981) will also be noted. For the problem of nonparametrically estimating a median, results analogous to those for finite population sampling observed in Section 2.8.2 will be given. In Section 3.7 we consider a situation where the labels of the units are such that members of the population whose labels are close together are more alike than units whose labels are far apart. For such a population the Polya posterior would not be appropriate since the necessary beliefs about exchangeability are not present. For such a case a sensible estimator of the population total is one that linearly interpolates between successive members of the sample. This estimator is shown to be admissible using the stepwise Bayes technique. The corresponding pseudo posterior is also shown to lead to sensible interval estimators.

3.1 Prior information

In this section we shall see some ways to generalize the arguments of Section 2.3.2 which allow for various types of prior information to be incorporated into the analysis. In the first part we consider situations where a prior guess for the population is available. In the second we consider situations where an auxiliary variable, $\mathbf{x}$, is present and our prior beliefs are exchangeable about certain functions of the y_i's and the x_i's. The results of this section were first presented in Vardeman and Meeden (1983). Although the results will be presented only for the problem of estimating the population mean similar admissibility results hold for the corresponding estimators of the population median.

3.1.1 A prior guess for the population

In this section we will assume that the statistician has a prior guess for the population, along with a measure of how much weight should be attached to the prior guess. Suppose the prior guess is a set of possible typical values, given by $\mathbf{a} = (a_1, \ldots, a_r)^T$, along with the corresponding probabilities $\mathbf{v} = (v_1, \ldots, v_r)^T$. This prior guess might reflect some fairly crude prior information about the centre or shape of the population. The length of $\mathbf{a}$, r, could be

considerably smaller than N, the population size, even as small as ten or five or even one. In particular a_i is not assumed to be related in any way with y_i. We let w be the weight we attach to our prior guess of the population. A large w reflects more confidence in our prior guess than a small w. Let $m = \sum_{i=1}^{r} a_i v_i$ denote the mean of our prior guess.

As we saw in (2.6), in the Bayesian paradigm with a prior π, given the observed data, $z = (s, z_s)$, one only needs to find $E_\pi(y_j|z)$ for each unobserved unit j to find the Bayes estimate of the population mean. The stepwise Bayes argument of Section 2.3.2 justified replacing this conditional expectation with the sample mean in situations with little or no prior information. This suggests that in the situation described here one could replace this conditional expectation by a convex combination of the sample mean and m, our prior guess for the mean. This reasoning leads one to the estimator

$$\delta_{m,w}(z) = N^{-1}\left\{\sum_{i\in s} z_i + (N - n(s))\left\{\frac{w}{w+n(s)}m + \frac{n(s)}{w+n(s)}\bar{z}\right\}\right\}. \tag{3.1}$$

Note that we have seen this estimator before in equation 1.14 where it arose from a two stage normal model. For $0 < w < \infty$ we see that the choice of w controls how heavily the guessed mean is to be weighted in comparison to the sample mean.

Two interesting special cases are $w = 0$ and $w = \infty$. In the first, $\delta_{m,0}$ is just the sample mean while in the second

$$\delta_{m,\infty} = N^{-1}\left\{\sum_{i\in s} z_i + (N - n(s))m\right\}.$$

Here the estimator assumes the seen and unseen are independent and just uses the prior guess to estimate each unseen unit. Among others, Godambe (1969) has studied the properties of this estimator.

We will now give a stepwise Bayes argument which establishes the admissibility of $\delta_{m,w}$ for certain finite parameter spaces. For a given vector of distinct real numbers $\mathbf{b} = (b_1, \ldots, b_k)^T$ and real number m we say that m is **compatible with b** provided there is a probability distribution on $\mathbf{b}$ with mean m.

Theorem 3.1 *If m is compatible with $\mathbf{b} = (b_1, \ldots, b_k)^T$, a vector*

of distinct real numbers, where m is a real number and w is a positive real number, then for estimating the population mean, under squared error loss, the estimator, $\delta_{m,w}$, is admissible, under any design p, when the parameter space is $\mathcal{Y}(\mathbf{b})$.

Proof. Let $\mathbf{v} = (v_1, \ldots, v_k)^T$ be a probability distribution over $\mathbf{b} = (b_1, \ldots, b_k)^T$ that can be used to establish the compatibility of m with $\mathbf{b}$. For convenience we will suppose the b_i's to be indexed so that $v_1, \ldots, v_r$ are positive and $v_{r+1}, \ldots, v_k$ are 0. Let B denote the set, $\{b_{r+1}, \ldots, b_k\}$.

We are now ready to define a partition of $\mathcal{Y}(\mathbf{b})$ which will serve as the regions of support for the successive mutually singular priors in the stepwise Bayes argument. For $j = 0, 1, \ldots, N$ let

$$A_j = \{\mathbf{y} \in \mathcal{Y}(\mathbf{b}) : \text{exactly } j \text{ different elements of } B \text{ are represented amongst } y_1, \ldots, y_N\}.$$

Now on the set A_j we define the prior distribution π^j by

$$\pi^j(\mathbf{y}) \propto \prod_{i=1}^{r} \Gamma(wv_i + c_y(i)) \prod_{\{i : i>r \text{ and } c_y(i)>0\}} \Gamma(c_y(i)).$$

Letting Λ^j be those samples which receive positive probability under π^j and the design p and 0 probability under π^l and p for $l < j$, it is straightforward to verify that $\delta_{m,w}$ is the stepwise Bayes estimator against this sequence of priors and hence is admissible by Theorem 2.1. In the proof we have assumed the set B was nonempty. If B is empty then $\delta_{m,w}$ is a Bayes estimator against the prior which is proportional to $\prod_{i=1}^{k} \Gamma(wv_i + c_y(i))$ and hence is admissible. □

In Section 3.6.2 we will discover a close relationship between the estimator, $\delta_{m,w}$, and Dirichlet process priors (Ferguson, 1973).

At first glance it may appear that the above argument was needlessly complicated by the introduction of the values of $\mathbf{a}$ along with the probabilities $\mathbf{v}$ as a prior guess for the population, since we only used the mean of this hypothesized population in constructing our estimator. Obviously it is easier to specify a mean rather than a whole distribution. However, when estimating parameters other than the population mean or when one is trying to find an interval estimate of a parameter we have seen that one wishes to simulate copies of the full population. As the above proof demonstrates this

can be done once $\mathbf{v}$ and $\mathbf{a}$ have been selected. We next consider an example to see how this approach could work in practice.

The population is a group of 332 large corporations which we denote by ppsales. The y variable is their total sales for the year 1975. Its mean is 2.41 and variance is 17.73 and as would be expected is strongly skewed to the right. The 10%, 30%, 50%, 70%, 90% and 95% quantiles are 0.57, 0.84, 1.24, 2.08, 4.79 and 6.91 respectively. The smallest value is 0.33 and the 12 largest values range from 9.96 to 44.86. The next largest value is 8.17. We also have an auxiliary variable, say x, which is their total sales for the year 1974. The correlation between the two variables is 0.997. ppsales is discussed further in Section 3.2 where the plot of y versus x is given in Figure 3.9. We will now use some of the information present in the auxiliary variable to select various $\mathbf{v}$'s as a prior guess for the population of y values. In each case we will take as our vector $\mathbf{a}$ the one which places equal probability on the members of $\mathbf{v}$.

For $\mathbf{v}_1$ we ordered the x values from largest to smallest and then took every sixth one starting with the third largest. The length of $\mathbf{v}_1$ was 55 with mean equal to 2.29, variance equal to 12.69 and median equal to 1.18. Hence this matches the y values quite well and represents better prior information than would be typically available.

We took $\mathbf{v}_2$ to be $(0.55, 0.76, 1.16, 2.26, 6.82)^T$ the 5%, 25%, 50%, 75% and 95% quantiles of the x values. Its mean is 2.31 and its variance is 6.80. It is a good guess for the centre of the population of y values even though it is under-dispersed. It can be thought of as a cruder version of $\mathbf{v}_1$ and in some situations might approximate the kinds of prior information that will be available. Simulations have indicated that selecting $\mathbf{v}$'s with as few as five values is not a good idea when one is interested in interval estimates. On the other hand we need not select nearly as many as 55 either. We will select four more $\mathbf{v}$'s that should more sensibly reflect what could happen in practice. That is, one needs to have some information about the centre of the population of interest and how dispersed it is. Often $\mathbf{v}$ can be assumed to be a guess for a set of representative quantiles.

We let $\mathbf{v}_3$ be $(0.51, 0.67, 0.83, 1.0, 1.16, 2.16, 2.66, 3.16)^T$ which has mean 1.53 and variance 0.88. Even though its median is equal to the median of $\mathbf{v}_2$ and reasonably close to the true median of 1.24 it is badly under-dispersed and its mean is too small.

We set $\mathbf{v}_4$ equal to $(0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.5, 2, 4, 6.5, 10)^T$. Its

mean is 2.65 and variance is 9.20. This is quite a good guess for the true population although its mean and median are farther away from the truth than those of $\mathbf{v}_1$.

For the next vector we shifted the lower half of our guess too far to the right so that we had a large over-estimate of the median. We set $\mathbf{v}_5 = (0.8, 1.05, 1.4, 1.75, 2, 2.5, 3.5, 6, 10)^T$ which has a mean of 3.22, a median of 1.75 and a variance of 8.89.

For the final vector we shifted the lower half of our guess too far to the left which resulted in an under-estimate of both the median and the mean. We set $\mathbf{v}_6 = (0.5, 0.6, 0.68, 0.75, 0.8, 0.95, 1.3, 3, 5)^T$ which has a mean of 1.51, a median of 0.8 and a variance of 2.30.

For each to these six vectors we took 500 simple random samples of size 30. We computed point and interval estimates for the population mean and median using 500 simulated copies of the entire population using the posterior of Theorem 3.1. We set the weight $w = 30$ so we are giving equal weight to the sample values and our prior guess. We compared these results to standard results. To get the confidence interval for the median we used the asymptotic version of Woodruff's method (Särndal *et al.*, 1992). The results are given in Table 3.1. We denote the standard Normal theory estimates by $NTmn$ and $NTmd$ for the mean and median respectively. The corresponding Polya posterior estimates are $gPmn_i$ and $gPmd_i$ where i denotes the fact that we are using the vector $\mathbf{v}_i$ as our prior guess. When estimating the mean we use squared error loss and when estimating the median we use absolute error loss.

The first thing we note from Table 3.1 is that the Normal theory interval for the mean under-covers because of the extreme skewness of the population. By considering the results for the first two vectors we see that when we have good prior information we can make significant gains over the standard frequentist procedures. Note that for these two cases the average error for estimating the mean is the same for the Polya posterior point estimates while the interval estimate is shorter for $\mathbf{v}_2$ than $\mathbf{v}_1$ and under-covers a bit. This shows the problem with selecting a $\mathbf{v}$ that is too short; it results in an interval estimate which is also too short. Probably the most surprising result in the table is that for estimating mean $gPmn_i$ has a smaller average error than $NTmn$ in all six cases. The interval estimate of the mean for the $gPmn_i$ is poorer in just two cases. Those are cases three and six where the prior guess for the mean is too small and the $\mathbf{v}_i$ is very under-dispersed. For estimating the median the point estimates based on $gPmd_i$ seem

Table 3.1 *Comparison of standard estimates and Polya posterior estimates which incorporate a prior guess for the population based on 500 simple random samples of size 30 from ppsales for six different prior guesses whose weight is equal to 30. We let lowbd denote the lower bound of an interval estimate.*

Estimator	Ave value	Ave error	Ave lowbd	Ave length	Freq. of coverage
$NTmn$	2.43	0.55	1.15	2.57	0.80
$gPmn_1$	2.37	0.17	1.69	1.83	0.99
$NTmd$	1.28	0.20	0.93	0.99	0.94
$gPmd_1$	1.26	0.10	0.98	0.66	0.99
$NTmn$	2.38	0.56	1.14	2.46	0.75
$gPmn_2$	2.35	0.17	1.74	1.51	0.90
$NTmd$	1.28	0.18	0.92	0.97	0.95
$gPmd_2$	1.24	0.10	0.95	0.81	0.95
$NTmn$	2.38	0.56	1.15	2.45	0.74
$gPmn_3$	1.99	0.34	1.53	1.24	0.55
$NTmd$	1.28	0.19	0.93	0.96	0.92
$gPmd_3$	1.27	0.13	0.96	0.78	0.96
$NTmn$	2.35	0.52	1.15	2.41	0.76
$gPmn_4$	2.49	0.16	1.83	1.61	0.97
$NTmd$	1.26	0.18	0.93	0.93	0.93
$gPmd_4$	1.23	0.11	0.97	0.68	0.97
$NTmn$	2.35	0.49	1.14	2.42	0.76
$gPmn_5$	2.75	0.26	2.08	1.62	0.91
$NTmd$	1.27	0.20	0.93	0.96	0.94
$gPmd_5$	1.65	0.41	1.25	0.86	0.48
$NTmn$	2.38	0.49	1.15	2.46	0.78
$gPmn_6$	1.98	0.33	1.48	1.33	0.58
$NTmd$	1.26	0.17	0.92	0.95	0.95
$gPmd_6$	1.03	0.21	0.81	0.52	0.84

a bit more sensitive to the $\mathbf{v}$'s than the mean estimates and the amount of improvement over the sample median is not as large as the improvement in the mean case. They perform poorly in cases five and six where there is a poor guess for the median. The median estimates do not seem to be so sensitive to $\mathbf{v}$ with too small variances. Taken together the results of the table suggest that the the posteriors of this section can be quite robust against certain kinds of misspecification. For cases like ppsales where the auxiliary variable is only partially and incompletely known there could often be enough prior information available to make a sensible choice of $\mathbf{v}$. It could be particularly useful for strongly skewed populations with a sample with few observations where standard methods for the mean work poorly.

3.1.2 A prior guess for each unit

The estimators of Section 3.1.1 treated the entries in $z = (s, z_s)$ in a symmetric manner. However, there are many situations where this would be inappropriate because of prior information, some of which could be of a quite limited nature, about the units. In such circumstances a semi-Bayesian statistician might be unwilling to try and elicit an entire prior distribution for $\mathbf{y}$ from a client but would be willing to obtain guesses for the values $y_1, \ldots, y_N$ from the consultee. Let us call those guessed values $\mathbf{x} = (x_1, \ldots, x_N)^T$. Another possible scenario, which could give rise to the x_i's, would be that they are the values of an auxiliary variable which are available to the statistician. In any case, no matter how they arose, we will assume that x_i is a sensible prior guess for y_i.

We will begin by assuming that the statistician's beliefs about the ratios $r_i = y_i/x_i$'s are roughly exchangeable. (Note for this case we are also assuming that each $x_i \neq 0$.) Such an assumption could be reasonable if the x_i's do not differ too much in size. It was suggested by Basu (1971) and Royall (1970) that for such a situation, after the data $z = (s, z_s)$ is observed, $\bar{r}_s = n(s)^{-1} \sum_{i \in s} (z_i/x_i)$ is a sensible estimate of y_j/x_j for each $j \notin s$. That is $\bar{r}_s x_j$ is a sensible estimate of y_j. This reasoning leads us to the estimator

$$\delta^r_{\mathbf{x}}(z) \quad = \quad N^{-1} \left\{ \sum_{i \in s} z_i + \bar{r}_s \sum_{j \notin s} x_j \right\} \tag{3.2}$$

as an estimator of the population mean. Meeden and Ghosh (1983)

proved the admissibility of this estimator when the parameter space was $\mathcal{R}^N$ by showing it was admissible for every finite parameter space where the ratios $r_i = y_i/x_i$'s can only assume a specified finite set of possible values. The argument was the one just used to prove Theorem 2.3 except that the sequence of priors was applied to the ratios, not to the y_i's. In the spirit of Section 3.1.1 one could modify this estimator by incorporating a prior guess, say m, for the mean of the ratios ($m = 1$ would in some sense seem the most natural). The resulting estimator would just be $\delta^r_{\mathbf{x}}$ with $(w/(w+n(s)))m + (n(s)/(w+n(s)))\bar{r}_s$ replacing $\bar{r}_s$ where w is the weight we attach to our prior guess m. In the same way, the admissibility of this new estimator follows from an obvious modification of the proof of the admissibility of $\delta_{m,w}$.

A useful way to think about the estimator $\delta^r_{\mathbf{x}}$ is that implicit in its use should be a prior belief that is (although not symmetric in $y_1, \ldots, y_N$) symmetric in $r_1, \ldots, r_N$. This line of thinking suggests yet more possibilities. For example, one might be willing to say that although their prior belief is not symmetric in $y_1, \ldots, y_N$, it is symmetric in the differences $d_i = y_i - x_i$. This leads to the estimator for the population mean

$$\begin{aligned} \delta^d_{\mathbf{x}}(z) &= N^{-1}\left\{\sum_{i\in s} z_i + \sum_{j\notin s} x_j + (N - n(s))\,\bar{d}_s\right\} \\ &= N^{-1}\left\{\sum_{i=1}^{N} x_i + N\left(\sum_{i\in s} z_i/n(s) - \sum_{i\in s} x_i/n(s)\right)\right\} \end{aligned} \tag{3.3}$$

which is just the usual difference estimator. (See e.g. page 99 of Raj (1968).)

Perhaps one can envision situations in which before declaring your beliefs to be symmetric, you might need to rescale the differences between the y_i's and x_i's according to some (known) constants $c_1, \ldots, c_N$. Then one can think of applying the above structure to the quantities $u_i = (y_i - x_i)/c_i$. This leads to the estimator of the population mean

$$N^{-1}\left\{\sum_{i\in s} z_i + \sum_{j\notin s} x_j + \bar{u}_s \sum_{j\notin s} c_j\right\}. \tag{3.4}$$

Of course, there are other possibilities as well. Every choice of N one-to-one functions $\psi_1, \ldots, \psi_N$ and the assumption of a symmetric prior belief about the quantities $u_i = \psi_i(y_i)$ leads to a new estimator of the population mean, given by

$$\delta_{\mathbf{x}}^{\psi}(z) = N^{-1}\left\{\sum_{i\in s} z_i + \sum_{j\notin s}\left\{n(s)^{-1}\sum_{i\in s}\psi_j^{-1}(u_i)\right\}\right\}, \qquad (3.5)$$

where ψ_j^{-1} is just the inverse function of ψ_j. From a practical point of view, however, very complicated choices of the functions ψ_i's would seem to presuppose a level of prior information and detailed analysis uncommon in finite population sampling problems of even moderate size. Nevertheless, these estimators of the population mean have the virtue of possessing stepwise Bayes derivations. Moreover they are relatively simple and make use of the kinds of prior information that can, in some cases, be available in a finite population sampling problem.

The finite parameter sets for which admissibility is easily established for the estimator $\delta_{\mathbf{x}}^{\psi}$ are those collections of $\mathbf{y}$ for which the u_i's can take only finitely many values, i.e., for a vector $\mathbf{b}$ of k distinct real numbers and a given set of N functions ψ we let

$$\begin{aligned}\mathcal{Y}^{\psi}(\mathbf{b}) \;=\; \{\mathbf{y} \;:\; &\text{such that for } i = 1, \ldots, N,\ \psi_i(y_i) = b_j \\ &\text{for some } j = 1, \ldots, k\}. \qquad (3.6)\end{aligned}$$

Then following the proof of Theorem 2.3 we have this admissibility result.

Theorem 3.2 *Let* $\mathbf{b} = (b_1, \ldots, b_k)^T$ *be a vector of distinct real numbers and* $\psi = (\psi_1, \ldots, \psi_N)^T$ *be a vector of one-to-one functions from R to R. Then for estimating the population mean, under squared error loss, the estimator* $\delta_{\mathbf{x}}^{\psi}$ *is admissible, under any design p, when the parameter space is* $\mathcal{Y}^{\psi}(\mathbf{b})$.

As we have seen with the estimator $\delta_{\mathbf{x}}^{r}$ we can combine the ideas of Section 3.1.1 with the approach given here to get an estimator of the population mean which assumes a prior belief that is exchangeable or symmetric in the ratios $r_i = y_i/x_i$'s and makes use of a prior guess for the mean of the ratios, along with a prior weight for the guessed mean. The same thing is true for the estimator $\delta_{\mathbf{x}}^{\psi}$. Suppose now that instead of estimating the population mean we wished to estimate another population parameter under

squared error loss. Just as in Section 3.1.1, we now suppose that we can make a prior guess about the population of $u_i = \psi_i(y_i)$ values, i.e. we can specify some set of typical values along with their respective probabilities. (Recall that this guessed population can be quite a bit smaller than the actual population of u_i's.) Finally, suppose we can assign a prior weight to our guessed population which reflects how good a guess we think it really is. Then, even though the analogous stepwise Bayes estimator cannot usually be computed in closed form, it will be admissible for the appropriately chosen finite parameter spaces.

3.2 Using an auxiliary variable

In Section 3.1.2 we considered the case where the statistician could make use of $\mathbf{x} = (x_1, \ldots, x_N)^T$ when making inferences about the population. We assumed that each x_i is nonzero and that $\mathbf{x}$ is completely known. The x_i's could be a Bayesian statistician's best guess for the y_i's or more traditionally a known auxiliary variable. In either case we assumed that x_i was a sensible prior guess for y_i. In (3.2) we considered the estimator, $\delta^r_{\mathbf{x}}$, which arose from the assumption that the ratios y_i/x_i's are roughly exchangeable. In Section 3.2.1 we will study other 'ratio' type estimators for this setup and show that they have a stepwise Bayes interpretation in the special case when the sample size is just two. In Section 3.2.2 we will consider the problem of estimating the population median in this setup and see that the pseudo posteriors that lead to $\delta^r_{\mathbf{x}}$ yield point and interval estimators of the population median which have good frequentist properties.

3.2.1 Ratio-type estimators

Let $\mathbf{x}$ be as in the above and $\mathbf{c} = (c_i, \ldots, c_N)^T$ be a vector of known positive real numbers. In this section we will consider ratio-type estimators given by

$$\delta^r_{\mathbf{x},\mathbf{c}}(z) = N^{-1}\left\{\sum_{i\in s} z_i + \left\{\sum_{i\in s}(y_i/x_i)\left(c_i/\sum_{l\in s} c_l\right)\right\}\sum_{j\notin s} x_j\right\} \tag{3.7}$$

for estimating the population mean. Before considering special cases of this estimator we note that it can be given a Bayes-like interpretation. After the data point $z = (s, z_s)$ has been observed, the ratios y_i/x_i's for $i \in s$ are known. Suppose now that one assumes for any unobserved ratio, y_j/x_j for $j \notin s$ that it takes on the value y_i/x_i with probability proportional to c_i for $i \in s$. Under this assumption for any $j \notin s$

$$E(y_j|z) = x_j \sum_{i \in s} \frac{y_i}{x_i} \frac{c_i}{\sum_{l \in s} c_l}$$

which then leads to (3.7). However, $\delta^r_{\mathbf{x},\mathbf{c}}$ is not typically a Bayes estimator since there may be no single prior generating it as the posterior expectation of the population mean. In this section we will investigate whether or not it is a stepwise Bayes estimator.

But first we consider some interesting special cases of $\delta^r_{\mathbf{x},\mathbf{c}}$. In the case $c_1 = \cdots = c_N$ it just becomes the estimator $\delta^r_{\mathbf{x}}$.

In the case $c_i = x_i (> 0)$, the resulting estimator is

$$N^{-1} \left\{ \left(\sum_{i \in s} y_i \right) \Big/ \left(\sum_{i \in s} x_i \right) \right\} \sum_{i=1}^{N} x_i,$$

the classical ratio estimator.

In the case $x_i > 0$ and $c_i = x_i^2$, the resulting estimator is

$$N^{-1} \left\{ \sum_{i \in s} y_i + \left\{ \left(\sum_{i \in s} y_i x_i \right) / \left(\sum_{i \in s} x_i^2 \right) \right\} \sum_{j \notin s} x_j \right\},$$

a 'regression-type' estimator.

In the case $x_i = p_i$ and $c_i = 1 - p_i$, where $0 < p_i < 1$ and $\sum_{i=1}^{N} p_i = n$ and the only samples with positive probability are those for which $n(s) = n$, the resulting estimator is

$$N^{-1} \sum_{i \in s} (y_i/p_i),$$

the Horvitz–Thompson estimator.

Again we suppose that $x_i = p_i$ and $c_i = 1 - p_i$ where $0 < p_i < 1$ and $\sum_{i=1}^{N} p_i = 1$. Now for a sample $s = (i_1, i_2)$ of size two the resulting estimator is

$$N^{-1}(2 - p_{i_1} - p_{i_2})^{-1} \left\{ \frac{y_{i_1}}{p_{i_1}}(1 - p_{i_2}) + \frac{y_{i_2}}{p_{i_2}}(1 - p_{i_1}) \right\},$$

the symmetrized Des Raj estimator based on PPSWOR samples

of size two where the p_i's are the selection probabilities of the first draw (Sengupta, 1980).

As these examples show, the estimator $\delta^r_{\mathbf{x},\mathbf{c}}$ includes several popular estimators as special cases and so the question of its admissibility is of some interest. In the next theorem we show that it is admissible in a very special case. But before stating the theorem we need some more notation. For $i = 1, \ldots, N$ let $\psi_i(y_i) = y_i/x_i$ be a function defined on the ith unit. Let ψ denote the collection of these functions and $\mathbf{b}$ be a vector of distinct real numbers. Then a parameter space for which the admissibility of $\delta^r_{\mathbf{x},\mathbf{c}}$ can be demonstrated is $\mathcal{Y}^{\psi}(\mathbf{b})$, which was defined in (3.6). This is the collection of those $\mathbf{y}$'s for which the ratios y_i/x_i can just take on a value b_j belonging to $\mathbf{b}$.

Theorem 3.3 *Let* $\mathbf{b} = (b_1, \ldots, b_k)^T$ *be a vector of distinct real numbers. Let* p *be a design such that* $p(s) = 0$ *if* $n(s) \neq 2$. *Then for estimating the population mean, under squared error loss, the estimator* $\delta^r_{\mathbf{x},\mathbf{c}}$ *is admissible when the parameter space is* $\mathcal{Y}^{\psi}(\mathbf{b})$.

Proof. The result will follow by showing that the estimator is in fact a stepwise Bayes estimator against a specific sequence of priors. The first prior puts mass k^{-1} on each of the k points $(b_1, \ldots, b_1)^T$, $\cdots, (b_k, \ldots, b_k)^T$. The only observed data points consistent with this prior are those where both of the observed ratios are the same and equal to one of the b_j's. For such a data point, where both the observed ratios are, say, b_j, the Bayes estimate under this prior is $b_j \sum_{i=1}^N x_i$ which agrees with $\delta^r_{\mathbf{x},\mathbf{c}}$ for such a data point.

In the next and final stage we will take care of all those sample points where the two observed ratios take on two different values. First we will take care of the data points where the two observed ratios have the values b_1 and b_2. To handle these points, for the restricted problem we define a prior which only puts positive mass on the set of $\mathbf{y}$'s where $N - 1$ of the ratios y_i/x_i are b_1 and the other is b_2, or just the reverse is true. Note that there are $2N$ such points. For such a point $\mathbf{y}$ let A_y be labels of the $N - 1$ units with the equal ratios. Then we take as our prior the one for which the amount of mass assigned to such a $\mathbf{y}$ is proportional to $\prod_{j \in A_y} c_j$. Let $z = (s, z_s)$ be a sample consisting of two units whose observed ratios are b_1 and b_2. Then it is easy to see that for any $j \notin s$ we have $P(y_j/x_j = b_1) = c_1/(c_1 + c_2)$ and so the stepwise Bayes estimate agrees with $\delta^r_{\mathbf{x},\mathbf{c}}$ in this case. Now the cases where the two

observed ratios are any other two distinct values can be handled in exactly the same way and so the proof is complete. Note the order in which we consider the different pairs of possible values is immaterial. □

We see that some of the pseudo posteriors that arise from the stepwise Bayes argument are perhaps a bit unusual. If both the observed ratios have the same value then the pseudo posterior assumes that the rest of the ratios must take on the same value. This is not surprising, since that is exactly what the Polya posterior does. But when the two observed ratios are different the pseudo posterior gives positive mass to only two possible parameter points. To be specific suppose we have observed $y_1/x_1 = b_1$ and $y_2/x_2 = b_2$ then the pseudo posterior puts mass $c_1/(c_1 + c_2)$ on the point $(b_1, b_2, b_1, \ldots, b_1)$ and mass $c_2/(c_1 + c_2)$ on the point $(b_1, b_2, b_2, \ldots, b_2)$. Note that this corresponds to a rather extreme prior belief about the composition of the population. Even so, having a stepwise Bayes justification for the estimator $\delta^r_{\mathbf{x},\mathbf{c}}$, in this special case, is informative because this yields what an analogous estimator would be when estimating another parameter. Without such information it is not at all clear how to find a comparable estimator for estimating the median, say.

For this reason it is unfortunate that we know of no stepwise Bayes justification for the estimator $\delta^r_{\mathbf{x},\mathbf{c}}$ other than in the special case covered in Theorem 3.3. In fact in Meeden and Ghosh (1981b) it was shown that for estimating the population mean when the parameter space is $\mathcal{Y}^{\psi}(\mathbf{b})$ the ratio estimator is inadmissible in the case where $N = 4$, $\mathbf{b} = (b_1, b_2)^T$ and p is such that $p(s) > 0$ if and only if $n(s) = 3$. This suggests that perhaps there is no general stepwise Bayes justification for the estimator $\delta^r_{\mathbf{x},\mathbf{c}}$ or, if there is, the finite parameter spaces need to be selected in a different manner. For the special cases of the ratio and Horvitz–Thompson estimators, their admissibility was demonstrated in Joshi (1965) and (1966), when the parameter space is $\mathcal{R}^N$.

In the special case of Theorem 3.3 Meeden and Ghosh (1981b) considered the uniform admissibility of the estimator $\delta^r_{\mathbf{x},\mathbf{c}}$ with respect to the class of designs of fixed sample size two. In various special cases they demonstrated that this estimator along with a deterministic design which always selects two units with the largest x_i values are uniformly admissibility. Uniform admissibility will be discussed further in Section 3.4.

3.2.2 Estimating the median

In the last section we considered the problem of estimating a population mean in the presence of an auxiliary variable. This problem has been widely discussed in the finite population sampling literature. The ratio estimator has often been used in such situations. For the problem of estimating a population median the situation is quite different. Only recently has this problem been discussed. Chambers and Dunstan (1986) proposed a simple method for estimating the population distribution function and the associated quantiles. They assumed that the value of the auxiliary variable was known for every unit in the population and their estimator came from a model-based approach. Rao *et al.* (1990) proposed ratio and difference estimators for the median using a design-based approach. Kuk and Mak (1989) proposed two other estimators for the population median. To use these estimators one only needs to know the values of the auxiliary variable for the units in the sample and its median for the whole population. The efficiencies of these estimators depend directly on the probability of 'concordance' rather than on the validity of an assumption of linearity between the variable of interest and the auxiliary variable. In this section we will compare some of these approaches with the approach based on the Polya posterior which assumes that one's beliefs about the ratios y_i/x_i are roughly exchangeable.

Some superpopulation models

As in Section 3.2.1 let $\mathbf{x} = (x_1, \ldots, x_N)^T$ be the known values of the auxiliary vector. Again we are assuming that x_i is a sensible guess for y_i. Before considering the problem of estimating the median we will briefly recall some facts about the superpopulation approach to estimating the mean. Consider the superpopulation model where it is assumed that for each i, $y_i = \beta x_i + u_i e_i$. Here β is an unknown parameter while the u_i's are known constants and the e_i's are independent identically distributed random variables with zero expectations. Since for a given sample z the population mean can be written as $N^{-1}(\sum_{i \in s} z_i + \sum_{j \notin s} y_j)$ we would expect $N^{-1}(\sum_{i \in s} z_i + \hat{\beta} \sum_{j \notin s} x_j)$ to be a sensible estimate of the mean whenever $\hat{\beta}$ is a sensible estimate of β. One particular choice of $\hat{\beta}$ is the weighted least squares estimator where the weights are determined by the u_i's. For example, if for all i, u_i is a constant

which does not depend on i then the resulting estimator is just the 'regression' through the origin estimator given in the last section. For example if for all i, $u_i = \sqrt{x_i}$, the resulting estimator is just the usual ratio estimator. While if for all i, $u_i = x_i$ the resulting estimator is $\delta^r_{\mathbf{x}}$ defined in (3.2). Note that in all three models the mean of the ratios y_i/x_i is always β. The variances, however, are different. In the first the variance of y_i/x_i is σ^2/x_i^2, while in the second it is σ^2/x_i and in the third it is σ^2 where σ^2 is the common variance of the e_i's. Note that in all three cases this agrees with the Bayes-like justification we gave earlier as special cases of (3.7). In particular, the constant variance in the third case is consistent with the stepwise Bayes justification for $\delta^r_{\mathbf{x}}$ which is appropriate when the y_i/x_i's are roughly exchangeable. Using this superpopulation setup it is easy to generate populations where each of the estimators has smaller mean squared error than the other two. A somewhat limited simulation study on a variety of populations found that the performances of the ratio estimator and $\delta^r_{\mathbf{x}}$ are quite similar, although in the majority of the cases the ratio estimator performs a bit better than $\delta^r_{\mathbf{x}}$. This is not unexpected, given the wide use of the ratio estimator.

We now turn to the problem of estimating the population median when an auxiliary variable $\mathbf{x} = (x_1, \ldots, x_N)^T$ is present and completely known. As we saw in Section 3.2.1, from the usual frequency point of view it is not clear how to extend the reasoning behind the ratio estimator for the mean to find an estimator of the median. This is not the case for the estimator $\delta^r_{\mathbf{x}}$, however. Note that under the assumption that the ratios y_i/x_i are roughly exchangeable one can, given an observed data point $z = (s, z_s)$, use the Polya posterior to simulate copies of the entire population of ratios. For such a simulated copy and $j \notin s$ suppose b_j is the simulated value of the ratio y_j/x_j. Then since for each j the value x_j is known, $b_j x_j$ is our simulated value for the unknown y_j. Therefore a simulated copy of the population of ratios leads in a natural way to a simulated copy of the population of y values with the observed units, $z(s) = \{ z_i : i \in s \}$, in place. For such a simulated copy of the y values one can find the median of the y_i's. If for the same z this process is repeated R times, where R is large, then we have approximated the predictive distribution of the population median under the Polya posterior. We then take the mean of this 'population' of simulated medians as our estimate of the population median. Even though the loss function will be absolute error,

just as in Section 2.8.2, we will use the mean of our 'population' of simulated medians rather than its median as our estimator. Let this estimator be denoted by estpp.

In what follows we will compare the estimator estpp to several other estimators. Another estimator we consider is just the sample median of z_s, the observed y of the units in the sample. This ignores the information contained in the auxiliary variable and is used as a benchmark. It will be denoted by estsm. Another estimator is the natural analogue of the ratio estimator of the population mean. This is discussed in Kuk and Mak (1989) and denoted by estrm. They propose two other estimators for the median. We will consider just the first one and denote it by estkm. Finally we will consider the estimator proposed in Chambers and Dunstan (1986) and denote it by estcd. We recall that estkm uses less information than do estpp and estcd since one only needs to know the values of the auxiliary variable in the sample and the median of the auxiliary variable for the entire population. Both the estimators estkm and estcd have plausible intuitive justifications. The argument for estcd is based on the superpopulation model described above with $u_i = \sqrt{x_i}$ for all i and its underlying spirit is not unlike the intuition leading to the Polya posterior.

Actually Chambers and Dunstan propose a whole family of estimators and we will only consider one special case which is appropriate when $u_i = \sqrt{x_i}$ in the superpopulation model described above. We now briefly outline the argument that leads to their estimator of the median. Let F denote the cumulative distribution function associated with the y values of the population. That is, F puts mass $1/N$ on each y_i in the entire population. The first step is to get an estimator of $F(t)$ for an arbitrary real number t. If s denotes our sample of size n then given the sample we can write

$$F(t) = N^{-1}\left\{\sum_{i\in s}\Delta(t-y_i) + \sum_{j\notin s}\Delta(t-y_j)\right\}$$

where $\Delta(u)$ is the step function which is one when $u \geq 0$ and zero elsewhere. Since the first sum in the above expression is known once we have observed the sample, to get an estimate of $F(t)$ it suffices to find an estimate of the second sum. Now under our assumed superpopulation model the population ratios $(y_i - bx_i)/\sqrt{x_i}$ are independent and identically random variables. Since after the sample s is observed a natural estimate of b is $\hat{b} = \sum_{i\in s} y_i / \sum_{i\in s} x_i$

one could act as if the n known ratios $(y_i - \hat{b}x_i)/\sqrt{x_i}$ for $i \in s$ are actual observations from this unknown distribution. Under this assumption, for a fixed t and a fixed unit j not in the sample s an estimate of $\Delta(t - y_j)$ is just the number of the n known ratios incorporating $\hat{b}$ less than or equal to $(t - \hat{b}x_j)/\sqrt{x_j}$ divided by n. Finally if we sum over all the unobserved units j these estimates of $\Delta(t - y_j)$ we then have an estimate for the second sum in the above expression for $F(t)$ which then yields an estimate of $F(t)$. Once we can estimate $F(t)$ for any t by, say, $\hat{F}(t)$ then the estimate of the population median is $\inf\{\, t \,:\, \hat{F}(t) \geq 0.5 \,\}$.

The populations

We will compare these estimators using six different artificial populations and three actual populations. For each of the nine populations we will plot y against x and y/x against x. The results can be seen in Figures 3.1 to 3.9.

For an artificial population the collection of values for the auxiliary variable $\mathbf{x}$ is chosen and then $\mathbf{y}$ is selected by specifying, for each i, the conditional distribution of y_i given x_i where all these conditional distributions are independent. This is exactly the same process used in Section 2.8.3 to generate bivariate populations when estimating the ratio of two medians. In fact, two of the populations considered here are just ppstskb and pplna with y_i' now being called x_i. In the population, ppg20a, the x_i's were a random sample from a gamma distribution with shape parameter 20 and scale parameter one. Then y_i was set equal to an observation from a normal population with mean $1.2x_i$ and variance x_i.

In the population, ppg5a, the x_i's were ten plus a random sample from a gamma distribution with shape parameter five and scale parameter one. Then y_i was set equal to an observation from a normal population with mean $3x_i$ and variance x_i.

In ppg5b the auxiliary variable is the same as in ppg5a while y_i was set equal to an observation from a normal population with mean $3x_i$ but with variance x_i^2.

In population ppexp2a, the auxiliary variable was 50 plus a random sample from the standard exponential distribution. Then y_i was set equal to 80 minus an observation from a normal population with mean x_i and variance $(0.6 \log x_i)^2$.

All the populations contain 500 units except ppstskb which has 1000. In most examples where ratio-type estimators are used both

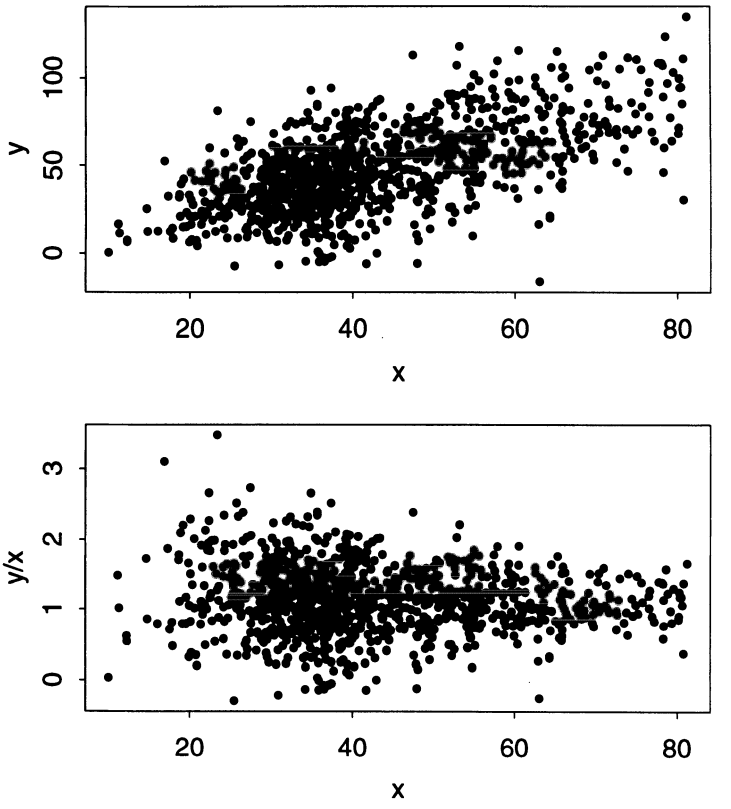

Figure 3.1 *For ppstskb the plot of y versus x and of y/x versus x.*

the y_i's and x_i's are usually strictly positive. This is true for all the populations here except for ppstskb which has 13 units with a negative y value.

Note that these populations were constructed under various scenarios for the relationship between the x and y variables. The populations ppg20a and ppg5a satisfy the assumptions of the superpopulation model leading to estcd, while ppg5b is consistent with the assumptions underlying estpp. In population ppstskb the conditional variance of y_i given x_i is consistent with estcd while for the unmodified pplna it was consistent with estpp. In both these cases the assumption for the conditional expectation is not sat-

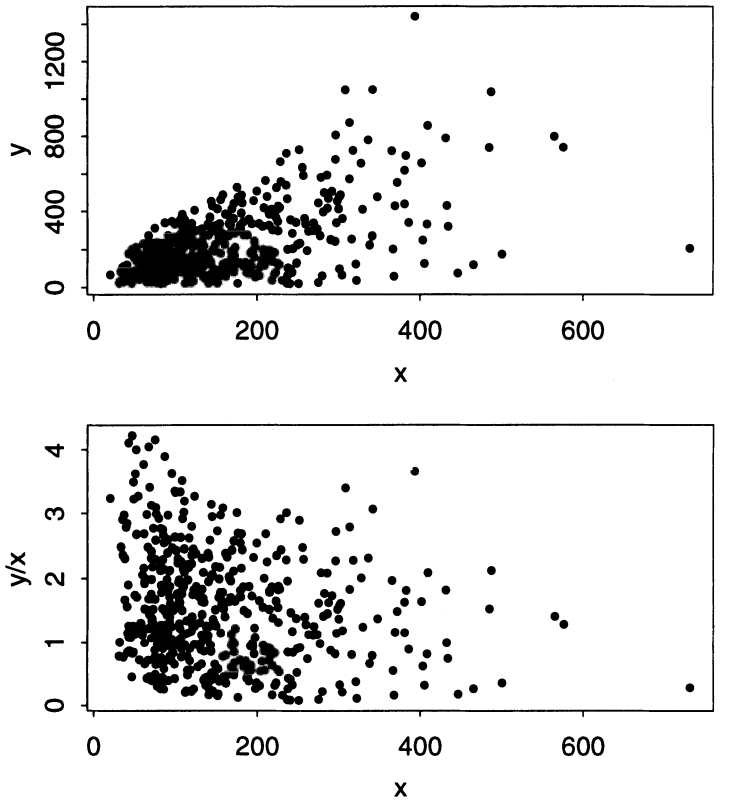

Figure 3.2 *For pplna the plot of y versus x and of y/x versus x.*

isfied. Finally, the population ppexp2a is an example where one would expect estcd and estpp to perform poorly. The correlation between the two variables for the six populations ppg20a, ppg5a, ppg5b, ppstskb, pplna and ppexp2a are 0.76, 0.87, 0.41, 0.61, 0.58 and -0.28 respectively.

The first of the three actual populations is a group of 125 American cities. The x variable is their 1960 populations, in millions, while their y variable is the corresponding 1970 populations, again in millions. The second is a group of 304 American counties. The x variable is the number of families in the counties in 1960, while the y variable is the total 1960 population of the county. Both

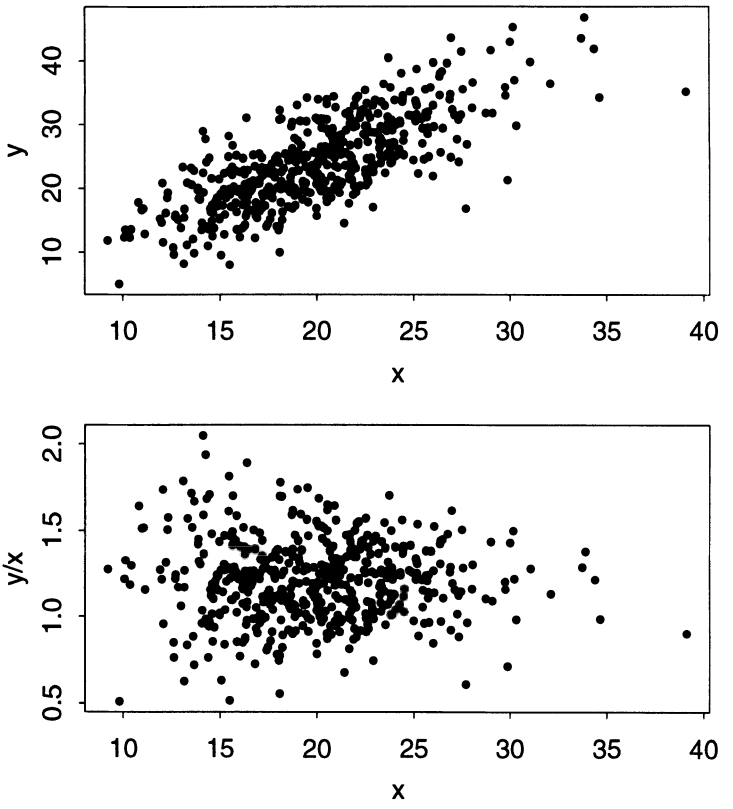

Figure 3.3 *For ppg20 the plot of y versus x and of y/x versus x.*

variables are given in thousands. The third population is 331 large corporations. The x variable is their total sales in 1974 and the y variable their total sales in 1975. The sales are given in billions of dollars. We denote these three populations by ppcities, ppcounties and ppsales. For the three populations the correlations are 0.947, 0.998 and 0.997. These populations were discussed in Royall and Cumberland (1981). Our ppcounties is similar to their population counties60 except that we have taken the x variable to be the number of families rather than the number of households.

As was noted earlier, for each of the nine populations we have plotted y against x and y/x against x in in Figures 3.1 to 3.9.

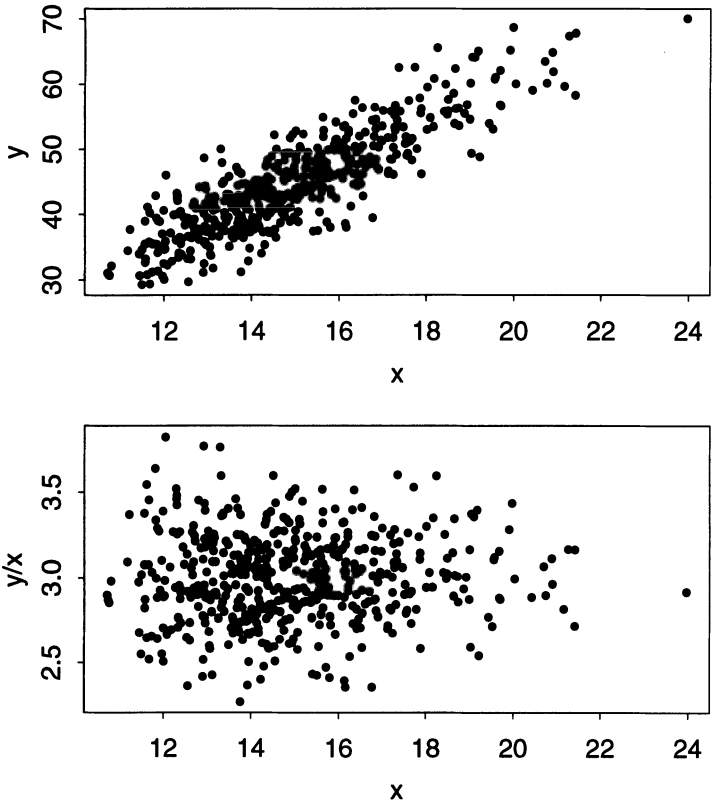

Figure 3.4 *For ppg5a the plot of y versus x and of y/x versus x.*

The estimator estpp is based on the assumption that given the sample s our beliefs about the observed ratios, i.e. the ratios y_i/x_i for $i \in s$ and the unobserved ratios, i.e. the ratios y_j/x_j for $j \notin s$ are roughly exchangeable. In particular this means that one's beliefs about a ratio y_j/x_j should not depend on the size of x_j. This is seen most clearly in the plot of the ratios for population ppg5b. On the other hand, under the superpopulation model leading to the estimator estcd we would expect the magnitude of the ratios to get smaller as the size of the x variable increases. This is seen clearly in the plot of the ratios for ppg20 and to a lesser extent for population ppg5a. In addition we see that for most of the remaining

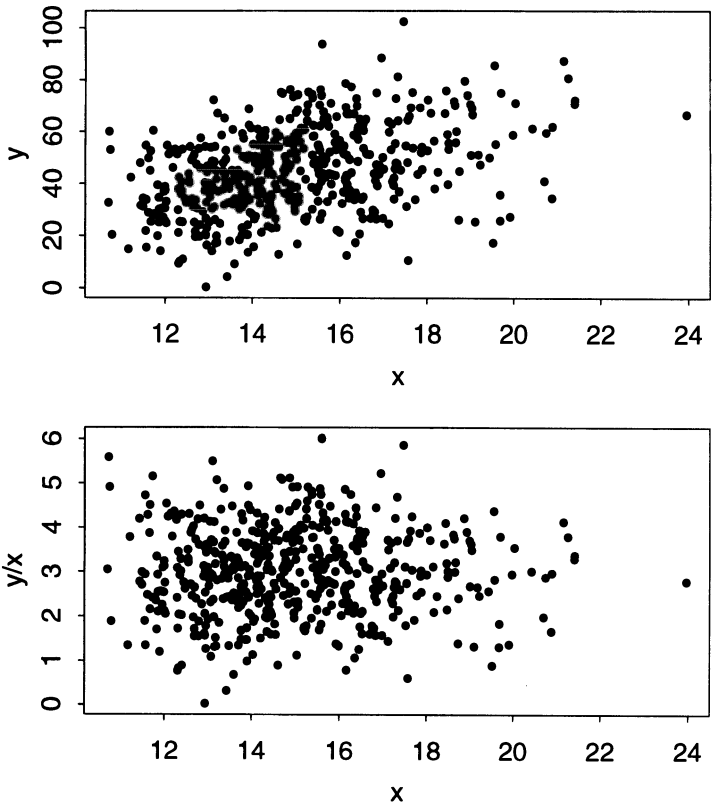

Figure 3.5 *For ppg5b the plot of y versus x and of y/x versus x.*

populations the size of a ratio does in fact depend on the size of the x variable in a variety of ways. Hence they should make interesting test cases for the estimator estpp.

Some simulation results

To compare the five estimators 500 simple random samples of various sizes were taken from the nine populations. For each case the average value and average absolute error of the estimator were computed. In each case estpp was found approximately using 500 simulations of the Polya posterior. In Table 3.2 on page 88 the average values of all the estimators except estsm are given. All the

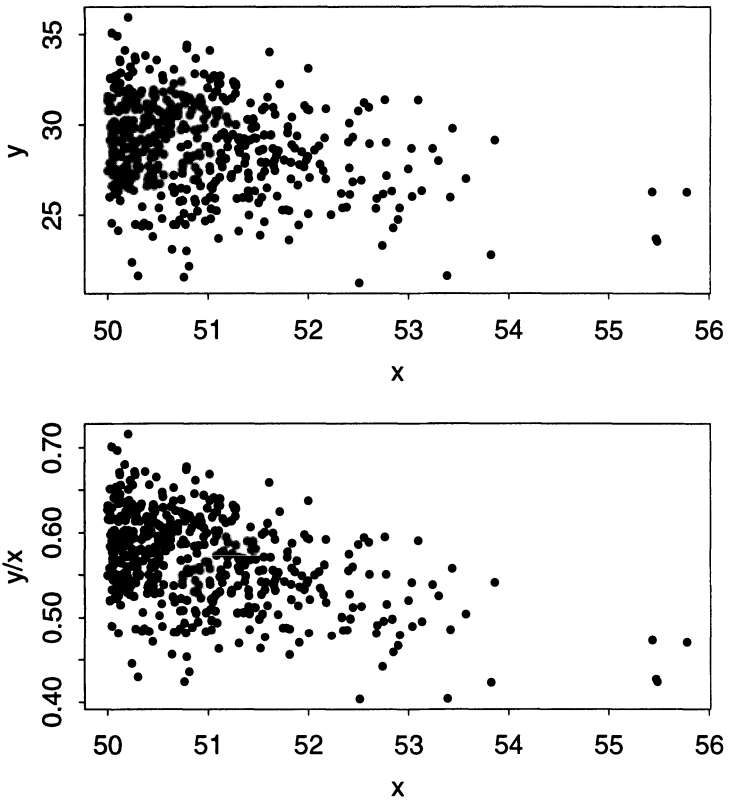

Figure 3.6 *For ppexp2a the plot of y versus x and of y/x versus x.*

estimators are approximately unbiased except in one case, estcd for the population pplna. In Table 3.3 on page 90 the average absolute errors for all five estimators are given. We see from this table that estcd and estpp are the clear winners. They both perform better than the other three estimators in every case but one. In ppexp2a they are both beaten by estsm, but this is one case where neither would be expected to do well. For the first seven populations their performances are nearly identical while for population pplna the estimator estpp is preferred and for population popstskb the opposite is true.

In practice one often desires interval estimates as well as point

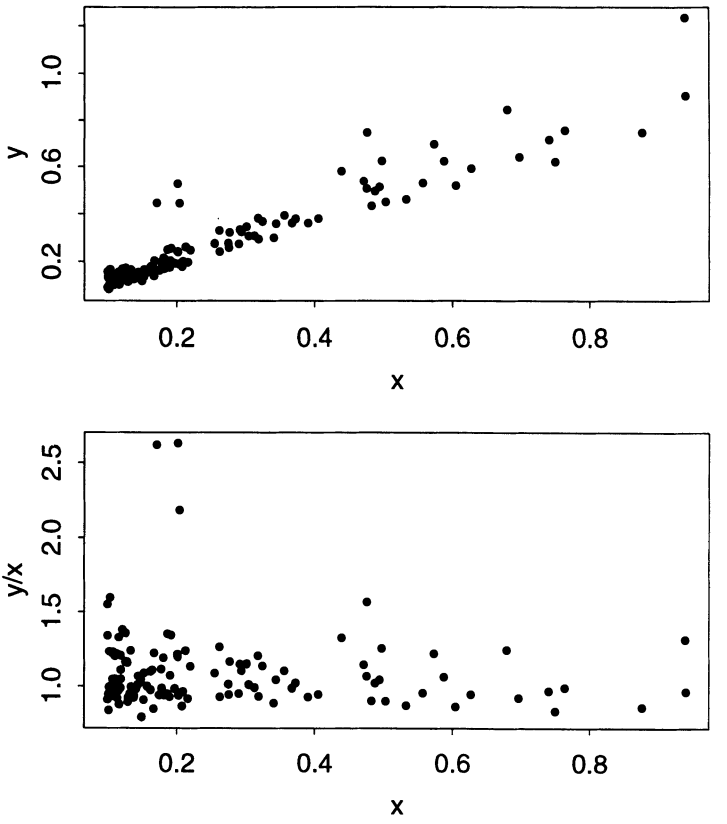

Figure 3.7 *For ppcities the plot of y versus x and of y/x versus x where x is the 1960 population (millions) and y is the 1970 population (millions) for 125 US cities.*

estimates for parameters of interest. Both Kuk and Mak (1989) and Chambers and Dunstan (1986) suggested possible techniques for finding interval estimates based on asymptotic theory; but in neither case did they actually implement their proposals. But as we have seen interval estimates based on the Polya posterior are easy to find. In Table 3.4 on page 92 we give the average length and frequency of coverage for approximate 0.95 credible intervals for the nine populations. Again we see that Polya posterior intervals have good frequentist properties.

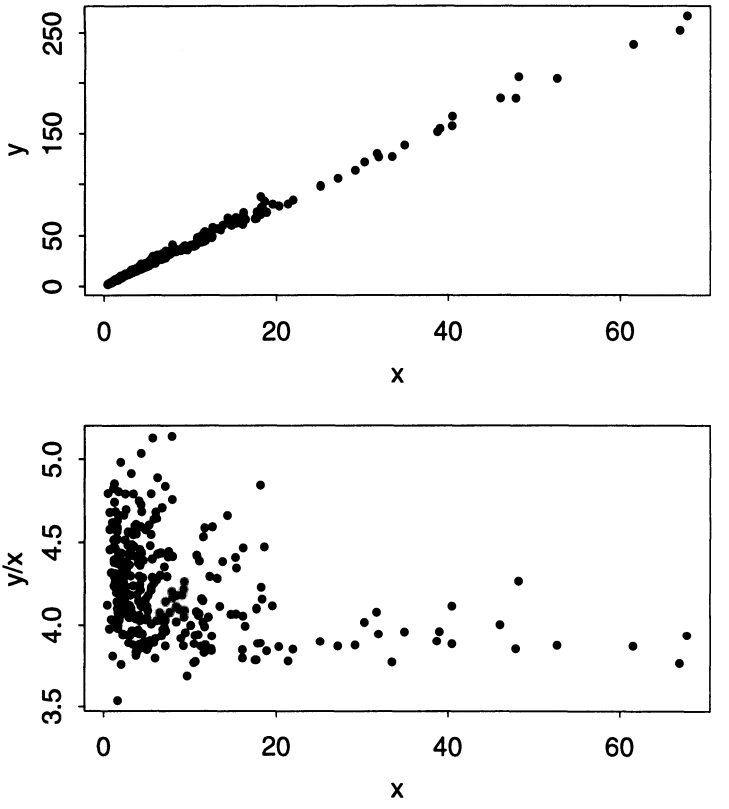

Figure 3.8 *For ppcounties the plot of y versus x and of y/x versus x where x is the number of families (thousands) living in a county and y is the total population (thousands) of the county for 304 counties.*

Discussion

The motivation for the estimator estpp is based on the assumption that the population ratios y_i/x_i are exchangeable. This assumption can be described mathematically in two separate but related ways. The first is the superpopulation model given earlier while the second comes from the Polya posterior which is based on a stepwise Bayes argument and gives a noninformative Bayesian interpretation for the estimator. This second approach is valid no matter what parameter is being estimated. When estimating the mean it

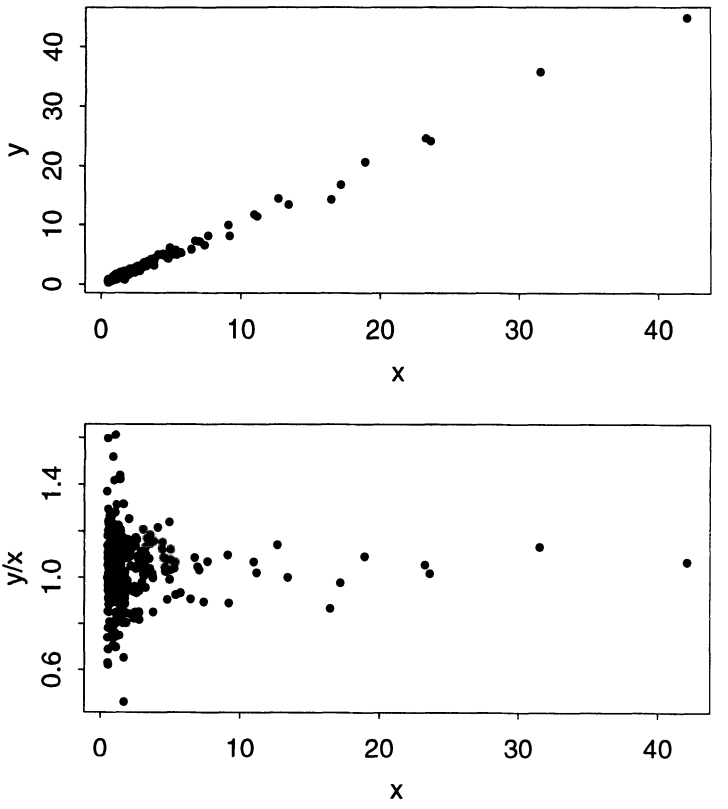

Figure 3.9 *For ppsales the plot of y versus x and of y/x versus x where x is the total 1974 sales (billions) and y is the total 1975 sales (billions) for 331 corporations.*

leads to the estimator given in (3.2) which performs very much like the ratio estimator, although the ratio estimator usually does a bit better. When estimating the median it leads to the estimator discussed in this section. Unfortunately there is no known Bayesian justification for the ratio estimator for the mean. If there were it could possibly lead to a sensible ratio estimator of the median as well. Here we have argued that the Polya posterior for the ratios leads to good point and interval estimators for the median when an auxiliary variable is present. Both the point and interval estima-

Table 3.2 *The average value of four estimators of the median for 500 simple random samples.*

Population (median)	Sample size	Average value of the estimator estrm	estkm	estcd	estpp
ppcities (0.19)	25	0.20	0.20	0.20	0.20
ppsales (1.24)	30	1.21	1.25	1.25	1.24
ppcounties (18.33)	30	18.21	18.60	18.26	18.39
ppexp2a (29.02)	30	29.03	29.05	29.03	29.05
ppg5a	30	43.82	43.88	43.99	43.89
(43.90)	50	43.90	43.91	44.06	43.90
ppg5b	30	43.84	43.96	44.15	43.61
(44.17)	50	44.28	44.37	44.18	43.98
ppg20a	30	23.47	23.28	23.46	23.77
(23.15)	50	23.34	23.18	23.43	23.18
pplna	30	171.15	169.38	185.01	170.61
(170.25)	50	169.15	167.54	185.03	169.61
ppstskb	30	43.66	40.27	45.50	45.11
(46.12)	50	44.04	40.70	45.43	45.37

tors are easy to find by simulating the Polya posterior and compare favourably to other methods. Moreover this approach seems to be reasonably robust against the assumption that the ratios y_i/x_i are exchangeable.

Royall and Cumberland (1981) gave an empirical study of the ratio estimator and estimators of its variance. They argued that

given a sample an estimate of variance based on the superpopulation model, which leads to the ratio estimator, often made more sense than a design-based estimate based on a probability sampling distribution. In Royall and Cumberland (1985), they demonstrated that, conditional on the sample mean of the auxiliary variable, the conditional coverage properties of the usual design-based confidence interval for the population mean were 'hopelessly unreliable'. This is consistent with the likelihood principle from which it follows that the design probabilities should play no role in the inferential process after the sample has been chosen. Note that inferences based on the Polya posterior do not depend on the design and hence are consistent with the likelihood principle. In the simulation studies done above simple random sampling was used for convenience. To get some idea of the conditional behaviour of the Polya posterior we considered five of our populations. In each case we ordered the population using the values of the auxiliary variable x. We then took 500 random samples from the first or smallest half of the population, then 500 more random samples from the second or largest half of the population and finally 500 more random samples from the middle third of the population. First we considered estimating the population mean and calculated the 95% confidence interval based on the ratio estimator and the 0.95 credible interval based on the Polya posterior which assumes the exchangeability of the ratios y_i/x_i. As was to be expected the relative frequency of coverage of both these intervals varied dramatically from a low of 0.02 to 1.0. Or, as Royall and Cumberland had noted, the conditional behavior of the ratio estimator, conditioned on the value of the average value of the auxiliary variable for the units appearing in the sample, can be quite different from its stated, nominal frequentist properties. The same is true for the Polya posterior interval for the mean. In Table 3.5 on page 93 we give the results for the Polya posterior estimators for the population median. (We also computed the average value and average absolute error of estcd for these examples. We did not include these results since they match closely the results of the Polya posterior.) We see that this approach performed quite well and its conditional behaviour, at least in these cases, is very much like its unconditional behaviour. In short, interval estimates for the median based on the Polya posterior should have reasonable frequentist properties, no matter how the sample was selected, as long as the population approximates our beliefs about the ratios are roughly exchangeable.

Table 3.3 *The average absolute error of five estimators of the median for 500 simple random samples.*

Population	Sample size	Average absolute error of the estimator estsm	estrm	estkm	estcd	estpp
ppcities	25	0.033	0.016	0.016	0.008	0.007
ppsales	30	0.180	0.077	0.080	0.024	0.025
ppcounties	30	3.12	0.59	0.96	0.22	0.21
ppexp2a	30	0.43	0.49	0.48	0.48	0.46
ppg5a	30	1.36	0.96	1.03	0.54	0.53
	50	0.95	0.74	0.78	0.44	0.43
ppg5b	30	2.84	2.74	2.71	2.37	2.38
	50	2.08	2.04	2.01	1.80	1.85
ppg20a	30	1.08	1.06	1.05	0.67	0.64
	50	0.94	0.77	0.78	0.51	0.49
pplna	30	25.9	25.8	24.2	21.4	17.0
	50	18.0	20.1	17.9	17.7	12.7
ppstskb	30	3.86	4.26	6.69	2.72	3.14
	50	2.92	3.63	5.82	2.20	2.51

As we can see by looking at the various figures and all our simulation results it does not seem to matter much if the variability in the ratios y_i/x_i decreases as x_i increases. What is crucial, however, is that the average value of the ratios in the narrow strip above a small interval of possible x values remains fairly constant as we move the small interval to the right. In Figure 3.5 on page 83 the plot of the ratios y_i/x_i for ppg5b is an example of what a plot should be like when we have exchangeability of the ratios. This is to be expected since that is how ppg5b was constructed. In Figure 3.4 on page 82 we see that the variability of the ratios decreases

a bit as we move to the right. Again this just reflects how ppg5a was constructed, a population consistent with the assumptions underlying the ratio estimator. Of the three real populations, ppsales is the one where the average value of the ratios in a narrow strip remains nearly constant as we move the strip to the right. For such populations most samples will yield a 'representative' set of ratios even though the ratios themselves are not exchangeable. For both pplna and ppstskb the average value of the ratios in a narrow strip tends to decrease as we move to the right. This reflects how the populations were constructed and explains the relatively poorer performance of the Polya posterior estimators in these cases, since many samples will yield a collection of 'unrepresentative' ratios.

All this suggests that when using the Polya posterior to make inferences one prefers a sample of units for which the observed ratios y_i/x_i are 'representative' of the population of ratios. Unfortunately, one will usually not be able to verify, in practice, if this is even approximately true. One possibility is to plot the observed ratios y_i/x_i against the x_i and observe whether or not the average value of the ratios in a narrow strip remains nearly constant as we move the strip to the right. If this is the case then the Polya posterior should yield interval estimators of the population median with good frequentist properties no matter how the samples are selected.

If the population ratios y_i/x_i really are exchangeable then it will not matter how the units are selected when we are using the Polya posterior to estimate either the mean or the median. Earlier we considered three very unbalanced sampling plans when estimating the median. As an alternative consider a more balanced sampling plan which is based on stratifying the population on the auxiliary variable. For example, consider again population pp5b, where the ratios are exchangeable, but where it has been ordered on the basis of its x_i values. We constructed ten strata where the first stratum consisted of the units with the 50 smallest x_i values, the second stratum of the units with the next 50 smallest x_i values and so on. We then took 500 stratified random samples of size 50 where five units were chosen at random from each stratum. For these samples the average value of estpp was 43.94 and its average absolute error was 1.81. The average length of its corresponding interval estimator was 8.95 with 0.938 relative frequency covering the true value. Note that these figures are very similar to those given earlier when simple random sampling was used.

Table 3.4 *The average length and relative frequency of coverage for a 0.95 credible interval for the median based on the Polya posterior for 500 simple random samples.*

Population	Sample size	Average length	Frequency of coverage
ppcities	25	0.041	0.968
ppsales	30	0.141	0.964
ppcounties	30	1.44	0.994
ppexp2a	30	2.26	0.944
ppg5a	30	2.70	0.950
	50	2.15	0.956
ppg5b	30	11.67	0.932
	50	8.86	0.942
ppg20a	30	3.24	0.960
	50	2.51	0.964
pplna	30	84.8	0.934
	50	65.4	0.956
ppstskb	30	15.52	0.936
	50	12.00	0.938

To summarize, these results indicate that the Polya posterior will lead to sensible inferences for the population median whenever the observed ratios are a 'representative' sample of the population ratios independently of how the sample was chosen. In addition, the Polya posterior leads naturally to both point and interval estimators. This is not the case for some of the other methods suggested for this problem where the interval estimation problem is more difficult.

Table 3.5 *The average value and absolute error for the point estimator and the average length and relative frequency of coverage for a 0.95 credible interval for the median based on the Polya posterior for 500 simple random samples of size 30, except for ppcities where it was 25, from the whole population, the 'smallest' half, the 'largest' half and the 'middle' third.*

Population	Where taken	Ave value	Ave error	Ave length	Freq. of coverage
ppcities	whole	0.195	0.0072	0.041	0.968
	small 1/2	0.192	0.0047	0.033	0.994
	large 1/2	0.196	0.0078	0.048	0.988
	middle 1/3	0.201	0.0114	0.055	0.922
ppcounties	whole	19.4	0.220	1.46	0.990
	small 1/2	18.6	0.305	1.34	0.942
	large 1/2	18.1	0.283	1.59	0.954
	middle 1/3	18.5	0.252	1.35	0.964
ppsales	whole	1.24	0.007	0.141	0.964
	small 1/2	1.24	0.027	0.153	0.966
	large 1/2	1.23	0.020	0.125	0.982
	middle 1/3	1.23	0.027	0.139	0.944
ppg5a	whole	43.9	0.53	2.70	0.950
	small 1/2	43.8	0.55	2.82	0.948
	large 1/2	44.0	0.53	2.55	0.940
	middle 1/3	43.9	0.47	2.63	0.974
ppg5b	whole	43.6	2.38	11.7	0.932
	small 1/2	42.2	2.69	11.6	0.890
	large 1/2	45.1	2.25	11.2	0.950
	middle 1/3	45.2	2.27	11.3	0.936

3.3 Stratification and prior information

Stratification has been one traditional method of incorporating prior information into finite population sampling problems. In this section we will provide Bayes and pseudo Bayes estimators which

use two different levels of prior knowledge about the stratification. In Section 3.3.1 we present the notation that we will use for these problems with stratification. In Section 3.3.2 we show how the approach developed in 3.1.1 can be extended to stratified populations when the prior information about stratum membership is vague. In Section 3.3.3 we present a generalization of this approach that will be useful in problems where one's prior information about stratum membership is more detailed. These results were first presented in Vardeman and Meeden (1984). Finally, in Section 3.3.4 we conclude with some examples.

3.3.1 Some notation

In addition to our usual notation of Section 1.1 we now need to be able to handle a stratified population. We assume that the population is stratified into strata $1, 2, \ldots, L$ where L is known. Attached to each unit i is a stratum membership h_i. The stratification may represent a division of the population into groups on the basis of the values of the y_i or on the basis of some other variable(s) possibly, but not necessarily, related to y_i. For a given $\mathbf{h}$ we let

$$str_k(\mathbf{h}) = \{\, i \,:\, h_i = k \,\}$$

be the units which belong to stratum k for $k = 1, \ldots, L$. (So k will denote the label of a typical stratum.)

In what follows we will make a variety of assumptions about one's knowledge concerning $\mathbf{h} = (h_1, \ldots, h_N)^T$. The case where $\mathbf{h}$ is completely known and is thus available for use in constructing estimators of μ, the population mean, is the situation of usual stratified sampling. In situations where the h_i are known only for those units sampled, it is common to use the term poststratification and the strata are sometimes called domains of study. See for example, Chapters 2 and 5A of Cochran (1977). We will use a framework that covers these two cases and others as well. We let $s = \{i_1, \ldots, i_{n(s)}\}$ and $s^* = \{j_1, \ldots, j_{n(s^*)}\}$ be two samples with $s \subset s^*$. A probability distribution, p, over such pairs (s, s^*) where s is not empty is a possible design. The intermediate case, where with p probability 1, $n(s) < n(s^*) < N$, is the case of double sampling for stratification, which is discussed in Chapter 12 of Cochran (1977). Our concern here is with estimators that incorporate prior information about the stratification. In any case a typical sample point will be $z = (s, s^*, z_s, z^*_{s^*})$ where $z^*_{s^*} = (z^*_{j_1}, \ldots, z^*_{j_{n(s^*)}})^T$, and

where $z^*_{j_l}$ is the stratum which contains unit j_l. That is for $i \in s^*$ we learn the stratum membership for the unit, while for $i \in s$ we also learn the value of the characteristic of interest as well. Note that for a given sample s we have that $str_k(\mathbf{h}(s))$ is the set of units in the sample which belong to stratum k. We let $n_k(\mathbf{h}(s)) = n_k$ denote the number of units that belong to this set. For the sample s^* we let n^*_k be the analogous quantity.

3.3.2 Prior information about stratum membership is vague

We begin by considering the situation where one's prior beliefs about the likely stratum memberships of all units are exchangeable. Suppose further that $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)^T$ are prior guesses at the relative sizes of the strata (with each $\eta_k \geq 0$ and $\sum_{k=1}^{L} \eta_k = 1$). Let $\mathbf{m} = (m_1, \ldots, m_L)^T$ be a prior guess for the means of the strata. Then for numbers w and $\mathbf{w} = (w_1, \ldots, w_L)^T$, all belonging to $[0, \infty]$, an estimator of the population mean incorporating these a priori values is

$$\delta_{\mathbf{m},w,\mathbf{w},\boldsymbol{\eta}}(z) = N^{-1}\left\{\sum_{i\in s} z_i + \sum_{k=1}^{L}\left(\left\{(n^*_k - n_k) + (N - n^*)\frac{w\eta_k + n^*_k}{w + n^*}\right\}\left\{\frac{w_k}{w_k + n_k}m_k + \frac{n_k}{w_k + n_k}\bar{z}_k\right\}\right)\right\} \tag{3.8}$$

where n_k and n^*_k are respectively the number of units in s and s^* belonging to stratum k, and $\bar{z}_k$ is the mean of those z_i's for units in s which belong to stratum k. When $n_k = 0$ we will understand $\bar{z}_k$ to be 0 and when both w_k and n_k are 0, we will take the factor

$$\hat{m}_k = \frac{w_k}{w_k + n_k}m_k + \frac{n_k}{w_k + n_k}\bar{z}_k$$

to be m_k.

To motivate this estimator, notice that $\hat{m}_k$ is a Bayes-like predictor of any unobserved unit known to belong to stratum k and that

$$\hat{\eta}_k = \frac{w\eta_k + n^*_k}{w + n^*}$$

is a Bayes-like estimator of the probability that a unit with $i \notin s^*$ belongs to stratum k. The estimator is then obtained by replacing each y_i having $i \in s^* - s$ with $\hat{m}_{h_i}$ and each y_i having $i \notin s^*$

with the $\hat{\eta}_k$ weighted average of the $\hat{m}_k$'s. The constants w and $\mathbf{w}$ of course control the relative weightings of the prior and sample values, ranging from domination of sample information in the '0' cases to domination of prior values in '∞' cases. The possibility that these constants are all 0 includes in our discussion the classical estimators common in stratified sampling, poststratification and double sampling for stratification. In this regard, compare the estimator with all w's equal to 0 to expressions on pages 91, 134 and 328 of Cochran (1977) respectively.

Note that this setup is a straightforward generalization of that presented in Section 3.1.1 to the stratified problem. This estimator has been studied in Binder (1982) in the instance that $n^* = N$. An interesting special case of this estimator was given in Hidiroglou and Srinath (1981) where p is such that $n = n^*$ and $L = 2$, $w = \infty$, $w_1 = 0$, $\eta_1 = 1$, $\eta_2 = 0$ and stratum 1 consists of exactly those units i with $y_i \leq C$, for C a large positive constant.

For cases of the estimator where w and each w_k and η_k are positive, it is possible to give proper Bayesian derivations for $\delta_{\mathbf{m},w,\mathbf{w},\boldsymbol{\eta}}$. This is also true in cases where w and/or some of the w_k's are ∞. Although this is no longer true when w or some of the w_k's are 0, the estimator then has a stepwise Bayes derivation. Hence a suitable modification of the argument of Section 3.1.1 will yield the admissibility of the estimator $\delta_{\mathbf{m},w,\mathbf{w},\boldsymbol{\eta}}$ for appropriately chosen finite parameter spaces. Moreover this estimator could be used by an only partially Bayesian sampler, armed with guesses of the relative sizes of the strata and strata means, and a notion of how to weight the various guesses against the information from the observed data.

Smith and Sedransk (1982) considered an interesting problem where a two-phased stratified sampling design is proposed. To estimate the age distribution of a fish population first a simple random sample is observed. The sampled fish are then stratified on the basis of their lengths and a further subsample is selected. The ages of these fish are determined by counting the annual growth rings on their scales. Since they are interested in estimating the age distribution of the fish population rather from the average age of the population their problem is somewhat different from the one we just discussed. Moreover they are primarily interested in determining optimal choices of the second-phase sample sizes, a problem we will not consider. However, the underlying probability structures of the two approaches are quite similar. In particular if only vague prior information is available about the sizes of the

strata based on the lengths of the fish then the 'posterior' for the '$w = 0$' case described above can also be used in their problem and be given a similar stepwise Bayes justification.

3.3.3 Prior information about stratum membership is less vague

In this section a generalization of the estimator $\delta_{\mathbf{m},w,\mathbf{w},\boldsymbol{\eta}}$ is presented that will be useful in problems where one's prior information about $\mathbf{h}$ is sharper than what was assumed in the previous section in that it is not exchangeable. As motivation, consider a situation where the strata are in fact defined in terms of the y_i themselves, with perhaps, 'low', 'medium' and 'high' strata having been defined in terms of the unknown values of the y_i. Suppose further that (perhaps on the basis of a census of y_i's at a previous period) the sampler can segment the N units into several groups indicating 'likely stratum membership'. In the present example, a three-group segmentation into 'likely low', 'likely medium' and 'likely high' seems natural, but a two-group segmentation into, say, 'likely low to medium' and 'likely medium to high' groups would also be possible, as would segmentations into more than three groups. In any case, although the sampler might well have the information needed to employ $\delta_{\mathbf{m},w,\mathbf{w},\boldsymbol{\eta}}$, it is an inappropriate estimator in that it ignores the prior information implicit in the ability to segment the population. We now define a generalization of $\delta_{\mathbf{m},w,\mathbf{w},\boldsymbol{\eta}}$ which would be appropriate in circumstances similar to these.

Suppose that in addition to y_i and h_i there is attached to unit i a variable g_i taking an integer value from 1 to G and suppose that $\mathbf{g} = (g_1, \ldots, g_N)^T$ is completely known. (In the example above, the values g_i would specify the subgroup memberships established by the sampler in the segmentation of the population using prior information and beliefs as opposed to the actual stratification of the population.) We let $q = 1, \ldots, G$ denote a typical value of the g_i's. Furthermore we will let N_{qk}, n^*_{qk} and n_{qk} be, respectively, the number of units in the population, s^* and s with $g_i = q$ and $h_i = k$. We will indicate sums over the subscripts with the usual dot notation. Then for each q we take $\boldsymbol{\eta}_q = (\eta_{q1}, \ldots, \eta_{qL})^T$ (with each $\eta_{qk} \geq 0$ and $\sum_{k=1}^{L} \eta_{qk} = 1$) as prior guesses of the relative sizes of $N_{q1}, \ldots, N_{qL}$. We let $\boldsymbol{\eta}$ denote the entire collection of the $\boldsymbol{\eta}_q$'s. As before let $\mathbf{m} = (m_1, \ldots, m_L)^T$ be guessed means for the strata.

Then for constants $\tilde{\mathbf{w}} = (\tilde{w}_1, \ldots, \tilde{w}_G)^T$ and $\mathbf{w} = (w_1, \ldots, w_L)^T$ our new estimator of the population mean is

$$\delta_{\mathbf{m},\mathbf{g},\tilde{\mathbf{w}},\mathbf{w},\boldsymbol{\eta}}(z)$$
$$= N^{-1}\Bigg\{\sum_{i\in s} z_i + \sum_{k=1}^{L}\Bigg(\Big\{(n^*_{\cdot k} - n_{\cdot k})$$
$$+ \sum_{q=1}^{G}(N_{q\cdot} - n^*_{q\cdot})\frac{\tilde{w}_q\eta_{qk} + n^*_{qk}}{\tilde{w}_q + n^*_{q\cdot}}\Big\}\Bigg)\Big\{\frac{w_k}{w_k + n_{\cdot k}}m_k + \frac{n_{\cdot k}}{w_k + n_{\cdot k}}\bar{z}_k\Big\}\Bigg\} \tag{3.9}$$

where as before, $\bar{z}_k$ is the mean of those units with $i \in s$ and $h_i = k$, we take $\bar{z}_k = 0$ when $n_k = 0$ and

$$\hat{m}_k = \frac{w_k}{w_k + n_{\cdot k}}m_k + \frac{n_{\cdot k}}{w_k + n_{\cdot k}}\bar{z}_k$$

is understood as m_k when both w_k and $n_{\cdot k}$ are 0 and

$$\hat{\eta}_{qk} = \frac{\tilde{w}_q\eta_{qk} + n^*_{qk}}{\tilde{w}_q + n^*_{q\cdot}}$$

is understood as η_{qk} when both $\tilde{w}_q$ and $n^*_{q\cdot}$ are 0.

This estimator can be obtained by replacing each y_i with $i \in s^* - s$ and $h_i = k$ by $\hat{m}_k$ and each y_i with $i \notin s^*$ with the $\hat{\eta}_{qk}$ weighted average of the $\hat{m}_k$'s. The w_k and $\tilde{w}_q$ govern how strongly the prior means and guessed stratum membership probabilities are weighted against the sample information.

An important special case of the above estimator is the 'all w's and $\tilde{w}$'s equal to 0' version given by

$$\delta_{\mathbf{g}}(z)$$
$$= \sum_{i\in s} z_i + \sum_{k=1}^{L}\Big\{(n^*_{\cdot k} - n_{\cdot k}) + \sum_{q=1}^{G}(N_{q\cdot} - n^*_{q\cdot})\frac{n^*_{qk}}{n^*_{q\cdot}}\Big\}\bar{z}_k, \tag{3.10}$$

provided each $n_{\cdot k} > 0$ and each $n^*_{q\cdot} > 0$. This is an attractive way to make use of the ability to stratify or segment the population and it uses no other prior information beyond our initial guesses for the strata given in $\mathbf{g}$.

Bayesian and stepwise Bayes justifications for these estimators can be made by modifying slightly the arguments of the previous section. The modifications necessary involve only specification of the marginal distribution of $\mathbf{h}$ (and not the conditional distribu-

tion of $\mathbf{y}$ given $\mathbf{h}$). The admissibility of the estimators $\delta_{\mathbf{m},w,\mathbf{w},\boldsymbol{\eta}}$ and $\delta_{\mathbf{m},\mathbf{g},\tilde{\mathbf{w}},\mathbf{w},\boldsymbol{\eta}}$ can be shown by generalizations of the basic stepwise Bayes argument. Unfortunately, many different cases must be considered and the notational burden, which is serve for $\delta_{\mathbf{m},w,\mathbf{w},\boldsymbol{\eta}}$, is even worse for $\delta_{\mathbf{m},\mathbf{g},\tilde{\mathbf{w}},\mathbf{w},\boldsymbol{\eta}}$. Here, to illustrate the type of argument that is required, without developing all the notation and different cases needed to prove a general result, we will outline a proof of the admissibility of the estimator $\delta_{\mathbf{g}}$, the estimator with all the $\tilde{w}_q$'s and w_k's set equal to zero. Recall that to prove the result, it suffices to specify a partition of the parameter space, a sequence of (mutually orthogonal) priors on the elements of this partition, and an ordering of this sequence and then show that the estimator in question is stepwise Bayes with respect to the ordered sequence.

Recall that $\mathbf{h} = (h_1, \ldots, h_N)^T$ where for each i, $1 \leq h_i \leq L$ and h_i is the stratum which contains unit i. We will assume that $\mathbf{h}$ is unknown and so a typical parameter point is now the pair $(\mathbf{y}, \mathbf{h})$. For $k = 1, \ldots, L$ we let $\mathbf{b}^k = (b_1^k, \ldots, b_{t_k}^k)^T$ be a vector of t_k distinct real numbers which are assumed to be the set of possible values for y_i for any unit in the kth stratum, i.e. for any unit i with $h_i = k$. Our parameter space is then given by

$$\mathcal{Y}(\mathbf{b}^1, \ldots, \mathbf{b}^L) = \{ (\mathbf{y}, \mathbf{h}) \ : \text{ such that for } i = 1, \ldots, N \\ y_i = b_j^k \text{ for some } j = 1, \ldots, t_k \text{ when } h_i = k \}.$$

For $k = 1, \ldots, L$ let

$$v_k(\mathbf{y}, \mathbf{h}) = \text{number of different values from } \mathbf{b}^k \\ \text{that appear for units } i \text{ with } h_i = k.$$

Recall that $\mathbf{g} = (g_1, \ldots, g_N)$ where for each i, $1 \leq g_i \leq G$ and g_i is the subgroup in which we have placed unit i. Hence $\mathbf{g}$ is our segmentation of the population based on our prior information and beliefs and is completely known. For $q = 1, \ldots, G$ let

$$u_q(\mathbf{h}) = \text{number of different values of } h_i \\ \text{that appear for units } i \text{ with } g_i = q.$$

Note that for any parameter point $(\mathbf{y}, \mathbf{h})$, that belongs to the parameter space defined above, we have that for each $k = 1, \ldots, L$ that $1 \leq v_k(\mathbf{y}, \mathbf{h}) \leq t_k$ and for each $q = 1, \ldots, G$ that $1 \leq u_q(\mathbf{h}) \leq L$. Let $\mathbf{v}(\mathbf{y}, \mathbf{h}) = (v_1(\mathbf{y}, \mathbf{h}), \ldots, v_L(\mathbf{y}, \mathbf{h}))^T$ and $\mathbf{u}(\mathbf{h}) = (u_1(\mathbf{h}), \ldots, u_G(\mathbf{h}))^T$.

Let $\mathbf{v}^o = (v_1^o, \ldots, v_L^o)^T$ be a specified vector of integer values which satisfy

$$1 \leq v_k^o \leq L \text{ for } k = 1, \ldots, L. \tag{3.11}$$

Let $\mathbf{u}^o = (u_1^o, \ldots, u_G^o)^T$ be a specified vector of integer values which satisfy

$$1 \leq u_q^o \leq G \text{ for } q = 1, \ldots, G. \tag{3.12}$$

For such a pair of vectors $\mathbf{u}^o$ and $\mathbf{v}^o$ we let

$$\mathcal{Y}(\mathbf{u}^o, \mathbf{v}^o) = \{ (\mathbf{y}, \mathbf{h}) \ : \ \mathbf{u}(\mathbf{h}) = \mathbf{u}^o \text{ and } \mathbf{v}(\mathbf{y}, \mathbf{h}) = \mathbf{v}^o \}$$

which is a (possibly void) subset of $\mathcal{Y}(\mathbf{b}^1, \ldots, \mathbf{b}^L)$. As $\mathbf{u}^o$ and $\mathbf{v}^o$ range over all possible choices that satisfy the constraints of equation (3.12) and equation (3.11), clearly the non-void $\mathcal{Y}(\mathbf{u}^o, \mathbf{v}^o)$ partition $\mathcal{Y}(\mathbf{b}^1, \ldots, \mathbf{b}^L)$. This gives the partition that will be used in the stepwise Bayes argument.

Next we must define a distribution on each non-void $\mathcal{Y}(\mathbf{u}^o, \mathbf{v}^o)$. We will do this by first giving a distribution for $\mathbf{h}$ and then a conditional distribution for $\mathbf{y}$ given $\mathbf{h}$. Let ϕ_q be the joint marginal distribution for the h_i's on the set with $g_i = q$. Then we take

$$\phi_q(h_i : g_i = q) \propto \prod_{k=1}^{L} \Gamma(N_{qk}).$$

(Notice that only u_q^o of the N_{qk} are positive and we will understand $\Gamma(0)$ to be 1.) We will assume that all these distributions are independent and hence the joint distribution of $\mathbf{h}$ is given by

$$\phi(\mathbf{h}) = \prod_{q=1}^{G} \phi_q(h_i : g_i = q).$$

To define the conditional distribution of $\mathbf{y}$ given $\mathbf{h}$ we will consider the sets of y_j's where $h_j = k$ for $k = 1, \ldots, L$ separately and then use independence again. For a given $k, \mathbf{y}$ and $\mathbf{h}$ let

$$c_y(i, k) \ = \ \text{number of } y_j\text{'s with } h_j = k \text{ and which equal } b_i^k,$$

then

$$\pi_h(y_j : h_j = k) \propto \prod_{i=1}^{t_k} \Gamma(c_y(i, k)),$$

where only v_k^o of the factors of this product are other than $\Gamma(0)$, and

$$\pi(\mathbf{y}|\mathbf{h}) = \prod_{k=1}^{L} \pi_k(y_j : h_j = k)$$

gives the conditional distribution of $\mathbf{y}$ given $\mathbf{h}$. Finally, we define as our distribution on the set $\mathcal{Y}(\mathbf{u}^o, \mathbf{v}^o)$ the prior

$$\pi_{\mathbf{u}^o,\mathbf{v}^o}(\mathbf{y}, \mathbf{h}) = \phi(\mathbf{h})\pi(\mathbf{y}|\mathbf{h}). \tag{3.13}$$

Next we need to specify an ordering of the non-void $\mathcal{Y}(\mathbf{u}^o, \mathbf{v}^o)$ and their corresponding priors. For this it suffices to place the $\mathcal{Y}(\mathbf{u}^o, \mathbf{v}^o)$ in lexicographical order according to their 'G+L digit labels' $(\mathbf{u}^o, \mathbf{v}^o)$. Then with $\Lambda^{(\mathbf{u}^o,\mathbf{v}^o)}$ standing for the set of sample points $(s, s^*, z_s, z^*_{s^*})$ possible under our design p and prior $\pi_{\mathbf{u}^o,\mathbf{v}^o}$, but not under p and any $\pi_{\mathbf{u}^1,\mathbf{v}^1}$ standing before it in the ordering, it is possible to verify that $\delta_{\mathbf{g}}$ is the $\pi_{\mathbf{u}^o,\mathbf{v}^o}$ conditional mean of the population mean given $(s, s^*, z_s, z^*_{s^*})$. This demonstrates that $\delta_{\mathbf{g}}$ is admissible for the parameter space $\mathcal{Y}(\mathbf{u}^o, \mathbf{v}^o)$.

3.3.4 Some examples

We conclude this section with some examples. We will only consider one special case of the stepwise Bayes model which yields the estimator $\delta_{\mathbf{g}}$. Recall that the population is divided into L strata defined by the vector $\mathbf{h}$. We will assume that $\mathbf{h}$ is completely unknown, so these true strata are not known. However, we do assume that the statistician has used prior information to select the vector $\mathbf{g}$ which contains a prior guess for the stratum membership of each unit in the population. For example if $g_i = k$ for an integer k between 1 and L then the statistician's best guess is that unit i belongs to stratum k, i.e. that $h_i = k$. As before we let N_{qk} be the number of units in the population with $g_i = q$ and $h_i = k$. Then $N_{1\cdot}, \ldots, N_{L\cdot}$ are sizes of guessed strata based on $\mathbf{g}$ and are known. On the other hand $N_{\cdot 1}, \ldots, N_{\cdot L}$ are the sizes of the true strata based on $\mathbf{h}$ and are unknown. We will also assume that the samples s and s^* are identical and so a typical sample point is $z = (s, z_s, z^*_s)$. Hence for the units in the sample z^*_s contains their true or actual stratum memberships. We let n_{qk} be the number of units in the sample with $g_i = q$ and $h_i = k$.

Now just as the vector $\mathbf{h}$ partitions the population into the L true strata the vector $\mathbf{g}$ partitions the population into the L guessed strata. In what follows we need some notation to denote the means and variances of these strata. We let $\mu^{\mathbf{g}}_q(\mathbf{y})$ and $\mathrm{Var}^{\mathbf{g}}_q(\mathbf{y})$ be the mean and variance of stratum q defined by $\mathbf{g}$ when the population is given by $\mathbf{y}$. We denote the corresponding sample quantities by $\bar{z}^{\mathbf{g}}_q$ and $\mathrm{Var}^{\mathbf{g}}_q(z)$. In a similar fashion we can denote for stratum

k defined by the vector $\mathbf{h}$ the mean and variance of the population and of the sample. In the usual way the population mean can be written as a weighted average of the strata population means defined by $\mathbf{g}$ or by $\mathbf{h}$, e.g. $\mu(\mathbf{y}) = \sum_{k=1}^{L}(N_{\cdot k}/N)\mu_k^{\mathbf{h}}(\mathbf{y})$ where N is the number of units in the population. This leads to the usual estimators which are just the corresponding weighted averages of the sample strata means. These estimators depend on the $N_{q\cdot}$'s and $N_{\cdot k}$'s respectively. Since the second quantities are unknown the usual estimator depending on $\mathbf{h}$ cannot be computed. Hence in standard frequentist practice the additional information about the true strata membership of the units in the sample, i.e. the information contained in z_s^* can not be used. The setup considered here is somewhat like the situation in standard practice where there is poststratification for groups or domains of study where the strata used in the design cut across these groups. In the above, $\mathbf{g}$ would define the strata and is known while $\mathbf{h}$ defines the groups and is only known for the elements in the sample. But it is often the case that the $N_{\cdot k}$'s are known a priori and hence the estimator based on $\mathbf{h}$ can be used.

Since we are assuming that the $N_{\cdot k}$'s are unknown the usual estimator based on the true strata defined by $\mathbf{h}$ cannot be used. However, as was seen in Section 3.3.3, in the proof of the admissibility of $\delta_{\mathbf{g}}$ the probability model underlying this estimator makes use of the information contained in z_s^* in a by now standard noninformative Bayesian manner. Conditioned on the sample $z = (s, z_s, z_s^*)$ there is a distribution for $\mathbf{h}$ and hence for the $N_{\cdot k}$'s as well. Then conditioned on the values of the $N_{\cdot k}$'s one just uses Polya sampling from the unseen to the seen within each true stratum k to get a simulated copy of the entire population where the Polya sampling is independent across the strata. We now find, under this model, the posterior variance of the population mean given the sample

$$
\begin{aligned}
&\mathrm{Var}(\mu(\mathbf{y})|z) \\
&\quad = \mathrm{Var}(E\{\mu(\mathbf{y})|z, N_{\cdot k}'s\}|z) + E(\mathrm{Var}\{\mu(\mathbf{y})|z, N_{\cdot k}'s\}|z) \\
&\quad = \mathrm{Var}\left(\sum_{k=1}^{L} \frac{N_{\cdot k}}{N} \bar{z}_k^{\mathbf{h}} \,\middle|\, z\right) \\
&\quad + E\left(\sum_{k=1}^{L} \left(\frac{N_{\cdot k}}{N}\right)^2 \frac{N_{\cdot k} - n_{\cdot k}}{N_{\cdot k}} \frac{\mathrm{Var}_k^{\mathbf{h}}(z)}{n_{\cdot k}} \frac{n_{\cdot k} - 1}{n_{\cdot k} + 1} \,\middle|\, z\right)
\end{aligned}
$$

Note that within each true stratum k we are using the expression

for the posterior variance under Polya sampling given in (2.15). In the special case when $L = 2$, i.e. we have just two strata, the first term in the above equation has a particularly simple form. Making use of the fact that $N_{\cdot 1} + N_{\cdot 2} = N$ we can write the above posterior variance as

$$\begin{aligned}\mathrm{Var}(\mu(\mathbf{y})|z) &= \mathrm{Var}(N_{\cdot 1}|z)\frac{(\bar{z}_1^{\mathbf{h}} - \bar{z}_2^{\mathbf{h}})^2}{N^2} \\ &+ N^{-2}\sum_{k=1}^{2} E\big(N_{\cdot k}[N_{\cdot k} - n_{\cdot k}] \mid z\big)\frac{\mathrm{Var}_k^{\mathbf{h}}(z)}{n_{\cdot k}}\frac{n_{\cdot k} - 1}{n_{\cdot k} + 1}\end{aligned} \tag{3.14}$$

Note that this variance depends on three different things. The first is the sample variances within the true strata. The second is the difference between the true strata sample means and the third is the posterior distribution of $N_{\cdot 1}$. Note that this in general will not be the same as the usual unbiased estimate of variance of the estimator based on $\mathbf{g}$ and simple random sampling within strata which is $N^{-2}\sum_{q=1}^{2} N_{q\cdot}(N_{q\cdot} - n_{q\cdot})\,\mathrm{Var}_q^{\mathbf{g}}(z)/n_{q\cdot}$. This is similar to the second term of the above equation with the $\mathrm{Var}_q^{\mathbf{g}}(z)$ terms replacing the $\mathrm{Var}_k^{\mathbf{h}}(z)$ terms and the known values $N_{q\cdot}$ replacing the expected values of $N_{\cdot k}$. If the true strata are more homogeneous than the guessed strata then one would expect that the second term of the above equation would be smaller than the estimated variance for the estimator based on the guess strata. However, the $\mathrm{Var}(\mu(\mathbf{y})|z)$ contains an additional term which increases its value. This term depends on the posterior variance of $N_{\cdot 1}$ and the difference between the true strata means in a sensible manner. We will now consider some examples where we compare the point and interval estimators of the population mean and median based on the stepwise Bayes posteriors from the model underlying the estimator $\delta_{\mathbf{g}}$ to the standard frequentist methods based just on the guessed strata defined by $\mathbf{g}$.

For simplicity we will just considered cases with two strata, i.e. $L = 2$. We constructed four different populations for our simulation study. The first was of size 500 consisting of two true strata defined by $\mathbf{h}$ of size 200 and 300 respectively. The first was a random sample from a gamma distribution with shape parameter four and the second was a random sample from a gamma distribution with

shape parameter seven and with three added to each value. The first true stratum had a mean of 3.87 and a variance of 3.40. The second true stratum had a mean of 10.05 and a variance of 6.56. The two guessed strata defined by **g** were each of size 250. The first guessed stratum contained 150 units from true stratum 1 and 100 units from true stratum 2. Consequently the second guessed stratum had 250 units from true stratum 2 and 50 units from true stratum 1. The mean and variance for guessed stratum 1 were 6.35 and 13.33 while the mean and variance for the guessed stratum 2 were 8.80 and 12.64. Note that the sizes of the guessed strata do not agree with the sizes of the true strata. We will denote this population by ppgst.

The next population was of size 400 where the first true stratum was a random sample of size 100 from a standard normal distribution and the second true stratum was a random sample of size 300 from a normal distribution of with mean two and variance one. The two guessed strata defined by **g** were each of size 200. The first guessed stratum contained 75 units from true stratum 1 and 125 units from true stratum 2. Consequently the second guessed stratum had 175 units from true stratum 2 and 25 units from true stratum 1. We will denote this population by ppnst2.

The last two populations were constructed as follows. Each have 400 units and the guessed strata and the true strata each contain 200 units. We began by taking a random sample of size 400 from the standard normal distribution. Half of these observations were selected to form true stratum 1. To form true stratum 2 we added three to all the units in the remaining half of the observations. To construct the first of these two populations we formed the guessed strata by placing half of true stratum 1 in guessed stratum 1 and the other half of true stratum 1 into guessed stratum 2. True stratum 2 was allocated in a similar fashion. We denote this population by ppnst3.50. Note that for this population the guessed strata are very inefficient since each is essentially a copy of the entire population. For the second of these populations the guessed strata will be more informative. It was constructed by placing three-fourths of true stratum 1 in guessed stratum 1 and the other fourth of true stratum 1 into guessed stratum 2. True stratum 2 was allocated in complementary fashion. We denote this population by ppnst3.75.

Using these four populations we compared standard frequentist methods based on the guessed strata defined by **g** for point and interval estimation of the population mean and median to those

based on the stepwise Bayes model underlying the estimator $\delta_{\mathbf{g}}$ given in Section 3.3.3. To get the frequentist confidence interval for the median we used the asymptotic version of Woodruff's method (Särndal *et al.*, 1992). To find the Polya posterior stepwise Bayes estimates approximately (only the estimate of the mean can be found explicitly) we used 500 simulated copies of the entire population for each random sample. The point estimate of the population median was the mean of the simulated medians. The 95% credible interval estimates were computed in the usual way by finding the 0.025 quantile and 0.975 quantile of the 500 simulated population means or medians. For each population we took 500 random samples for three different sample sizes. For the first 500 samples we selected at random without replacement independent samples of size 10 from each of the two guessed strata. For the second the samples were of size 20 within each strata and for the third the samples were of size 40 within each strata. Some of the results are given in Table 3.6. We denote the standard normal theory estimates by $NTmn$ and $NTmd$ for the mean and median respectively. The corresponding stepwise Bayes estimates are $\mathbf{g}mn$ and $\mathbf{g}md$. We did not include the results for the point and interval estimates for the mean since the behaviour of the two approaches were very, very similar. The only difference occurred in the sample size 10 case where the credible interval for $\mathbf{g}mn$ tends to be a bit shorter on average than the standard interval and consequently undercovers a bit as well. We have seen this behaviour before. The factor $(n_{\cdot k} - 1)/(n_{\cdot k+1})$ is decreasing the posterior variance by a nontrivial amount as the sample size $n_{\cdot k}$ becomes small. We also omitted the results for the point estimates for the median for the cases with sample sizes 20 and 40 because again the behaviour of the two methods was quite similar. Table 3.6 contains the results for the case where the sample size was 10.

We see from the table that just as when estimating the mean with a small sample size the 0.95 credible intervals for the median are on the average a bit too short to cover the true population median 95% of the time. On the other hand the point estimator for $\mathbf{g}md$ always performs better than that for $NTmd$ since it is not as biased. This is true for the larger sample sizes as well but the improvement is not as large. This is another example of the phenomenon first noted in Section 2.8.2 that the Polya posterior can yield improved point estimators of the population median over standard frequentist methods.

Table 3.6 *Comparison of the standard 95% interval estimator and 0.95 credible intervals based on a stepwise Bayes model defined by a vector* $\mathbf{g}$ *for the population median. The results are based on sets of 500 random samples where the sample size within each stratum was 10. The error is the absolute value of the difference between the estimate and the population median.*

Pop. (median)	Est.	Ave value	Ave error	Ave length	Freq. of coverage
ppgst	$NTmd$	8.05	0.94	4.99	0.932
(7.88)	$\mathbf{g}md$	7.68	0.83	4.49	0.930
ppnst2	$NTmd$	1.82	0.25	1.48	0.952
(1.76)	$\mathbf{g}md$	1.72	0.24	1.29	0.936
ppnst3.50	$NTmd$	0.71	0.23	1.20	0.972
(0.65)	$\mathbf{g}md$	0.62	0.21	1.10	0.940
ppnst3.75	$NTmd$	1.81	0.60	2.44	0.958
(1.50)	$\mathbf{g}md$	1.64	0.47	2.26	0.908

Perhaps it is somewhat surprising that the estimators based on the stepwise Bayes model which yields the estimator $\delta_{\mathbf{g}}$ only do better than standard methods in the problem of finding a point estimator of the population median given that they are using the additional information of true stratum membership contained in z_s^*. To help see why this is so, consider the special case where $L = G = 2$ and each member in true stratum 1 all take on the same value, say a, and each member in true stratum 2 all take on a different value, say 0. In this case the second term in (3.14) is zero since $\mathrm{Var}_k^{\mathbf{h}}(z)$ is zero for $k = 1$ and $k = 2$ and $\mathrm{Var}(\mu(\mathbf{y})|z)$ is just the product of a^2 and $\mathrm{Var}(N_{\cdot 1}|z)$. Under the stepwise Bayes model giving $\delta_{\mathbf{g}}$ and assuming that $N_{1\cdot}$ and $N_{2\cdot}$ are large the posterior distribution of $N_{q1}/N_{q\cdot}$ given z is approximately Beta with parameters n_{q1} and $n_{q\cdot} - n_{q1}$. Since under this model the posterior distributions of these two ratios are independent it follows that

$$\mathrm{Var}(\mu(\mathbf{y})|z) \doteq \frac{a^2}{N^2} \sum_{q=1}^{2} N_{q\cdot}^2 \frac{n_{q1}(n_{q\cdot} - n_{q1})}{n_{q\cdot}^2(n_{q\cdot} + 1)}$$

Now it is easy to check that for this special case the estimated variance of the usual stratified estimator based on the guessed strata is just the above expression with the finite population correction factor for a stratum, i.e. $1 - n_{q\cdot}/N_{q\cdot}$, included as a factor in the term in the summation. So even though the $\mathrm{Var}_k^{\mathbf{h}}(z)$'s should be smaller than the $\mathrm{Var}_k^{\mathbf{g}}(z)$'s the first term on the right-hand side of (3.14) involving $\mathrm{Var}(N_{\cdot 1})$ also contributes significantly to $\mathrm{Var}(\mu(\mathbf{y})|z)$. One way to decrease this contribution is to include a prior guess η_q for the relative sizes of $N_{q1}, \ldots, N_{qL}$ for each guessed stratum q and $\tilde{\mathbf{w}}$ a vector of constants used to weight our prior guess for stratum membership. This is a special case of the estimator given in (3.9) with $\mathbf{m}$ and $\mathbf{w}$ set equal to a vector of zeros.

To see how this could work we present the results for a small simulation study for ppnst3.75. Recall in this population 75% of true stratum 1 belongs to guessed stratum 1 and the remaining 25% of true stratum 1 belongs to guessed stratum 2. We considered the case where a sample of size 20 was taken from each guessed stratum. We selected $\tilde{\mathbf{w}}$ so that our prior guess for stratum membership was given equal weight with the sample information contained in z_s^*. We consider three different choices of the vectors η_q's. For the first case we took as our case the truth, i.e. we guessed that 75% of guessed stratum 1 belonged to true stratum 1 and that 25% of guessed stratum 2 belonged to true stratum 1. For the other two cases our guesses were 90% and 10% respectively and 90% and 40% respectively. Note that in the first of these two, even though we overestimate in guessed stratum 1 and underestimate in guessed stratum 2, overall we still have a corrected estimate of the sizes of the two true stratas. This, however, is not true in the second case where we overestimate the size of stratum 1 in both guessed strata and hence overestimate the size of true stratum 1. When estimating the mean we used square error loss and absolute error loss when estimating the median. The results of the simulations are given in Table 3.7.

As expected the results show that correct prior information can yield improved procedures. Moreover we see that getting the prior guess for the relative sizes of the true strata within each guessed stratum is not nearly important as having a good overall guess of the true strata sizes. Simulations for two cases with prior guesses $\eta = (0.60, 0.40)$ and $\eta = (0.60, 0.10)$ reconfirmed this fact. The results for the first case are very similiar to the results for the $\eta =$

Table 3.7 *Comparison of standard point estimators and 95% interval estimators and point estimators and 0.95 credible intervals based on a stepwise Bayes model defined by a vector* $\mathbf{g}$ *and* $\eta = (\eta_{11}, \eta_{21})$, *a prior guess for the relative sizes of the two true strata within each guessed stratum. The results are based on sets of 500 random samples from population ppnst3.75 where the sample size within each stratum was 20.*

Est.	$\eta = (\eta_{11}, \eta_{21})$	Ave value	Ave error	Ave length	Freq. of coverage
NTmn		1.58	0.060	0.95	0.936
g*mn*	(0.75,0.25)	1.57	0.031	0.81	0.972
NTmd		1.51	0.405	1.81	0.956
g*md*	(0.75,0.25)	1.59	0.259	1.70	0.988
NTmn		1.58	0.063	0.95	0.926
g*mn*	(0.90,0.10)	1.58	0.035	0.75	0.936
NTmd		1.49	0.423	1.83	0.958
g*md*	(0.90,0.10)	1.59	0.276	1.60	0.962
NTmn		1.59	0.065	0.95	0.946
g*mn*	(0.90,0.40)	1.39	0.073	0.81	0.854
NTmd		1.52	0.426	1.82	0.928
g*md*	(0.90,0.40)	1.18	0.377	1.59	0.840

(0.90, 0.10) case while the results for the second are very similiar to the $\eta = (0.90, 0.40)$.

In summary these results suggest that the stepwise Bayes model underlying $\delta_{\mathbf{g}}$ and more generally the model underlying the estimator in (3.9) incorporate the information of the true stratum membership of the units in the sample in a sensible way. A procedure based on $\mathbf{g}$ and prior guesses η still would not be as good as a poststratified procedure based on the assumed knowledge of the $N_{\cdot k}$'s, the group or domain sizes defined by $\mathbf{h}$. However, it could be used when one is not willing to assume that the $N_{\cdot k}$'s are known but one still has enough prior information for the prior guesses η and a choice of $\tilde{\mathbf{w}}$.

3.4 Choosing between experiments

Consider again the basic finite population sampling problem described in Section 1.1 with parameter space $\mathcal{Y}$ and loss function $L(\gamma(y), \cdot)$ for estimating $\gamma(y)$. In many such problems we must not only choose our estimator but also must choose our design p from some class of possible designs Φ. Two such classes which are often considered for a given positive integer n $(n < N)$ are $\Phi^1 = \{p : \sum_{s \in S} n(s)p(s) \leq n\}$, the class of designs of expected sample size less than or equal to n and $\Phi^2 = \{p : p(s) = 0 \text{ if } n(s) \neq n\}$, the class of designs of fixed sample size n. A pair (p, δ) is said to be **uniformly admissible** relative to the class Φ if $p \in \Phi$ and if there does not exist any other pair (p', δ') with $p' \in \Phi$ which dominates (p, δ). In this section we will see how the stepwise Bayes technique can identify such uniformly admissible pairs. In Section 3.4.1 we consider a more general mathematical formulation for choosing between experiments and in Section 3.4.2 we specialize it to the finite population sampling problem.

3.4.1 A more general problem

Consider the estimation problem where y, the true but unknown state of nature, is known to belong to some finite set $\mathcal{Y}$. We wish to estimate some real valued function $\gamma(y)$ with strictly convex loss function $L(\gamma(y), \cdot)$ defined on the decision space $\mathcal{A}$, a closed and bounded interval of real numbers. Before making the decision however, we may choose, possibly at random, to observe one of k different experiments. Let $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_k)^T$ be such that $\nu_i \geq 0$ and $\sum_{i=1}^k \nu_i = 1$. $\boldsymbol{\nu}$ is called a design and if we use $\boldsymbol{\nu}$ then we observe the ith experiment with probability ν_i. Our problem is to choose a design $\boldsymbol{\nu}$ and then a decision rule for each possible experiment.

Let $Z_1, \ldots, Z_k$ be the finite sample spaces of the k (≥ 2) experiments available to us. For $i = 1, \ldots, k$ and $y \in \mathcal{Y}$ let $f_{i,y}(\cdot)$ be the probability function on Z_i when y is the true state of nature. For each i and $z \in Z_i$ we assume that there exists a $y \in \mathcal{Y}$ such that $f_{i,y}(z) > 0$. Finally, let δ_i denote a typical estimator (possibly randomized) from Z_i to $\mathcal{A}$ with risk function $R_i(y, \delta_i)$.

Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)^T$. For the decision maker an estimation procedure for this problem is a pair $(\boldsymbol{\nu}, \boldsymbol{\delta})$. For such a pair its risk

function is

$$R(y, \boldsymbol{\delta}, \boldsymbol{\nu}) = \sum_{i=1}^{k} \nu_i R_i(y, \delta_i).$$

Let Φ be the set of possible designs which are available to the decision maker. Then a pair $(\boldsymbol{\nu}, \boldsymbol{\delta})$ is uniformly admissible relative to Φ if $\boldsymbol{\nu} \in \Phi$ and if there does not exist another pair $(\boldsymbol{\nu}', \delta')$ with $\boldsymbol{\nu}' \in \Phi$ and $R(y, \delta', \boldsymbol{\nu}') \leq R(y, \boldsymbol{\delta}, \boldsymbol{\nu})$ for all $y \in \mathcal{Y}$ with strict inequality for at least one y. Note if $\boldsymbol{\nu}$ is such that $\nu_i = 0$ for some i, then we will only consider pairs $(\boldsymbol{\nu}, \boldsymbol{\delta})$ where the corresponding member of $\boldsymbol{\delta}$ is unspecified, i.e., for a given design there is no need to consider estimators for experiments which are impossible to observe.

If π is a prior distribution over $\mathcal{Y}$ let $q_i(\cdot; \pi)$ be the marginal probability function on Z_i under π, then

$$r(\boldsymbol{\delta}, \boldsymbol{\nu}, \pi) = \sum_{i=1}^{k} \nu_i r_i(\delta_i, \pi)$$

is the Bayes risk of the pair $(\boldsymbol{\nu}, \boldsymbol{\delta})$ against the prior π where $r_i(\delta_i, \pi)$ is the Bayes risk of δ_i against π.

Let $\pi^1, \ldots, \pi^m$ be a sequence of mutually orthogonal prior distributions over $\mathcal{Y}$, i.e. $(\pi^i, \pi^j) = \sum_l \pi_l^i \pi_l^j = 0$ for $i \neq j$. We wish to define what it means for $\boldsymbol{\delta}$ to be stepwise Bayes against this sequence. Before doing this we introduce some more notation following that of Theorem 2.1. For $i = 1, \ldots, k$ let

$$\Lambda_i^1 = \{ z \,:\, z \in Z_i \text{ and } q_i(z; \pi^1) > 0 \}$$

and for $j = 2, \ldots, m$,

$$\Lambda_i^j = \left\{ z \,:\, z \in Z_j,\ z \notin \bigcup_{r=1}^{j-1} \Lambda_i^r \text{ and } q_i(z; \pi^j) > 0 \right\}.$$

Furthermore we assume that for $j = 1, \ldots, m$ that $\bigcup_{i=1}^{k} \Lambda_i^j$ is nonempty. In such a case we say that $\boldsymbol{\delta}$ is **stepwise Bayes** against $\pi^1, \ldots, \pi^m$ if for $i = 1, \ldots, k$ there exist $1 \leq r_1 < r_2 < \cdots < r_t \leq m$ depending on i such that $\bigcup_{u=1}^{t} \Lambda_i^{r_u} = Z_i$ and δ_i is stepwise Bayes against $\pi^{r_1}, \pi^{r_2}, \ldots, \pi^{r_t}$. That is each δ_i is stepwise Bayes against a subsequence of the π^j's.

Given a set of estimators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)^T$ which is stepwise Bayes against the sequence $\pi^1, \ldots, \pi^m$ we want to find a design $\boldsymbol{\nu}$

such that the pair $(\boldsymbol{\nu}, \boldsymbol{\delta})$ is uniformly admissible relative to Φ. The following theorem helps to simplify this problem.

Theorem 3.4 *Let Φ be the class of possible designs for the estimation problem with finite parameter space $\mathcal{Y}$. Let $\pi^1, \ldots, \pi^m$ be a sequence of mutually orthogonal prior distributions and suppose that the vector of estimators $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k)^T$ is stepwise Bayes against this sequence. For $w = 1, \ldots, W$ define the following subsets of Φ:*

$$\Phi_w = \left\{ \boldsymbol{\nu} \in \Phi_{w-1} \ : \ r(\boldsymbol{\delta}, \boldsymbol{\nu}, \pi^w) = \inf_{\boldsymbol{\nu}' \in \Phi_{w-1}} r(\boldsymbol{\delta}, \boldsymbol{\nu}', \pi^w) \right\}$$

where $\Phi_0 = \Phi$ and $W \leq m$. Then for any $\boldsymbol{\nu} \in \Phi_W$ the pair $(\boldsymbol{\nu}, \boldsymbol{\delta})$ is admissible relative to Φ if and only if $(\boldsymbol{\nu}, \boldsymbol{\delta})$ is admissible relative to Φ_W.

Proof. We will only show that if a pair is admissible relative to Φ_W then it is admissible relative to Φ, since the other direction is trivially true. So assume $(\boldsymbol{\nu}, \boldsymbol{\delta})$ is admissible relative to Φ_W but there is a pair $(\boldsymbol{\nu}', \boldsymbol{\delta}')$ which dominates it for the problem with Φ as the class of possible designs. From this it follows that

$$r(\boldsymbol{\nu}', \boldsymbol{\delta}', \pi^1) \leq r(\boldsymbol{\nu}, \boldsymbol{\delta}, \pi^1). \tag{3.15}$$

If Λ_i^1 is not empty then $r_i(\delta_i, \pi^1) \leq r_i(\delta', \pi^1)$ and so

$$r(\boldsymbol{\nu}', \boldsymbol{\delta}, \pi^1) \leq r(\boldsymbol{\nu}', \boldsymbol{\delta}', \pi^1). \tag{3.16}$$

Now if for some i there exists a $z \in Z_i$ such that $\delta_i(z) \neq \delta_i'(z)$ then equation (3.16) is strict and so

$$r(\boldsymbol{\nu}', \boldsymbol{\delta}, \pi^1) < r(\boldsymbol{\nu}, \boldsymbol{\delta}, \pi^1)$$

which contradicts the fact that $\boldsymbol{\nu} \in \Phi_1$. Hence for $i = 1, \ldots, k$ we have that $\delta_i(z) = \delta_i'(z)$ for $z \in \Lambda_i^1$ and so $\boldsymbol{\nu}' \in \Phi_1$ as well.

Using exactly the same argument we can show that $\delta_i(z) = \delta_i'(z)$ for $z \in \Lambda_i^2$ and that $\boldsymbol{\nu}' \in \Phi_2$. Repeating this argument $m - 2$ more times we see that $\boldsymbol{\delta}$ and $\boldsymbol{\delta}'$ are identical and that $\boldsymbol{\nu}' \in \Phi_W$. But this is not possible since by assumption $(\boldsymbol{\nu}, \boldsymbol{\delta})$ is admissible relative to Φ_W. □

It is easy to give examples where $\boldsymbol{\delta}$ is stepwise Bayes against the sequence of priors $\pi^1, \ldots, \pi^m$ and $\boldsymbol{\nu} \in \Phi_W$ but the pair is not admissible relative to Φ (Mazloum and Meeden, 1987). One useful condition that guarantees that every member of Φ_W yields an admissible pair is that $\cup_{i=1}^m \mathcal{Y}(\pi^i) = \mathcal{Y}$ where $\mathcal{Y}(\pi^i)$ is the support of π^i. This is easy to check and the proof will be omitted.

In the next section we will see how this theorem can be used in finite population sampling.

3.4.2 A finite population sampling example

We will now apply the results of the previous section to a finite population sampling problem. In finite population sampling the various experiments are just the different samples $s \in S$ which can be observed. A design is just the usual sampling design p. Let $\mathbf{x} = (x_1, \ldots, x_N)^T$ be a vector of auxiliary values which are available to the statistician. We assume that each of the x_i's is a positive constant and they are not all equal. In addition we assume that the statistician's prior beliefs about the ratios $r_i = y_i/x_i$'s are roughly exchangeable. For such a case we proved the admissibility of the estimator $\delta^r_{\mathbf{x}}$ in Section 3.1.2. The parameter space used in the proof was $\mathcal{Y}^{\psi}(\mathbf{b})$ defined in (3.6) in the special case where $\psi_i(y_i) = y_i/x_i$. We will denote this parameter space by $\mathcal{Y}^{\boldsymbol{r}}(\mathbf{b})$.

We now consider the problem of finding a design p such that the pair $(p, \delta^r_{\mathbf{x}})$ is uniformly admissible when the class of possible designs is $\Phi^1 = \{p : \sum_{s\in S} n(s)p(s) \leq n\}$ where n is a positive integer such that $n < N$. For $k = 1, \ldots, N$ let

$$S_k(max) = \left\{ s : n(s) = k \text{ and } \sum_{i\in s} x_i = \max_{s':n(s')=k} \sum_{i\in s'} x_i \right\}.$$

As we see in the next theorem a design which leads to a uniformly admissible pair is one which concentrates its mass on members of $S_n(max)$.

Theorem 3.5 *Consider the problem of estimating the population mean with squared error loss. Let* $\mathbf{b} = (b_1, \ldots, b_k)^T$ *be a vector of distinct real numbers. If* p *is a design such that* $\sum_{s\in S_n(max)} p(s) = 1$, *then the pair* $(p, \delta^r_{\mathbf{x}})$ *is uniformly admissible relative to* Φ^1 *when the parameter space is* $\mathcal{Y}^{\boldsymbol{r}}(\mathbf{b})$ *and hence is uniformly admissible when the parameter space is* $\mathcal{R}^N$.

Proof. For a given design p and the estimator $\delta^r_{\mathbf{x}}$ it is easy to see that

$$R(y, \delta^r_{\mathbf{x}}, p) = N^{-2} \sum_{s\in S} \left\{ \sum_{i\in S} \left(\sum_{j\notin S} x_j/n(s) \right) r_i - N^{-2} \sum_{i\notin S} x_i r_i \right\}^2 p(s)$$

where $r_i = y_i/x_i$ for $i = 1, \ldots, N$. For a given s and $i \in s$ we let $a_{i,s} = \sum_{j \notin s} x_j/n(s)$ while for $i \notin s$ we let $a_{i,s} = -x_i$. Hence

$$R(y, \delta^r_{\mathbf{x}}, p) = N^{-2} \sum_{s \in S} \left\{ \sum_{i=1}^{N} a_{i,s} r_i \right\}^2 p(s).$$

Let π^i be a prior in the sequence against which $\delta^r_{\mathbf{x}}$ is stepwise Bayes. Since under this prior the ratios $r_i = y_i/x_i$'s are exchangeable we have that

$$\begin{aligned} & r(\delta^r_{\mathbf{x}}, p, \pi^i) \\ & \quad = N^{-2} \left\{ E(r_1^2) \sum_{s \in S} \left(\sum_{j=1}^{N} a_{j,s}^2 \right) p(s) \right. \\ & \quad \left. + E(r_1 r_2) \sum_{s \in S} \left(\sum_{i \neq j} a_{i,s} a_{j,s} \right) p(s) \right\}. \end{aligned}$$

Since for each s, $\sum_{i \neq j} a_{i,s} a_{j,s} = -\sum_{i=1}^{N} a_{i,s}^2$ this equation becomes

$$r(\delta^r_{\mathbf{x}}, p, \pi^i) = N^{-2} \left\{ E(r_1^2) - E(r_1 r_2) \right\} \sum_{s \in S} \left(\sum_{i=1}^{N} a_{i,s}^2 \right) p(s).$$

By the Schwarz inequality $E(r_1 r_2) \leq (Er_1^2)^{1/2} (Er_2^2)^{1/2} = Er_1^2$ and a design belonging to Φ^1 will belong to Φ^1_W if for each $w = 1, \ldots, W$ it attains the following infimum:

$$\begin{aligned} & \inf_{p \in \Phi^1} r(\delta^r_{\mathbf{x}}, p, \pi^w) \\ & \quad = N^{-2} \left\{ E(r_1^2) - E(r_1 r_2) \right\} \inf_{p \in \Phi^1} \sum_{s \in S} \left(\sum_{i=1}^{N} a_{i,s}^2 \right) p(s) \\ & \quad = N^{-2} \left\{ E(r_1^2) - E(r_1 r_2) \right\} \\ & \quad \times \inf_{p \in \Phi^1} \sum_{s \in S} \left\{ \left(\sum_{i \notin s} x_i \right)^2 / n(s) + \sum_{i \notin s} x_i^2 \right\} p(s). \end{aligned}$$

Let $S(max) = \cup_{k=1}^{N} S_k(max)$ and $\Phi^1(max)$ be the subset of Φ consisting of all those designs which concentrate all their mass on samples belonging to $S(max)$. Clearly any design which attains the infimum of the above equation must belong to $\Phi^1(max)$.

We now assume that the population is labelled so that $x_1 \geq$

$x_2 \geq \cdots \geq x_N$. Therefore

$$\inf_{p \in \Phi^1} r(\delta^T_{\mathbf{x}}, p, \pi^w)$$

$$= N^{-2}\{ E(r_1^2) - E(r_1 r_2) \} \times \inf_{p \in \Phi^1(max)} \sum_{i=1}^{N} \psi(i) p(i)$$

where, with a slight abuse of notation, $p(i)$ is the probability, under design p, of selecting a sample of size i which contains the ith largest x_i's and

$$\psi(i) = i^{-1} \left(\sum_{j=i+1}^{N} x_j \right)^2 + \sum_{j=i+1}^{N} x_j^2$$

for $i = 1, \ldots, N$. Note $\psi(N) = 0$.

Let $\tilde{\psi}$ be the function defined on the interval $[1, N]$ which is obtained from ψ by connecting the points $(i, \psi(i))$ and $(i+1, \psi(i+1))$ with straight line segments for $i = 1, \ldots, N-1$. The proof of the theorem will be essentially completed once we prove that $\tilde{\psi}$ is a strictly decreasing convex function on $[1, N]$. It is easy to check that $\tilde{\psi}$ is strictly decreasing. We now show that it is convex as well.

Let i_0 be an integer satisfying $2 < i_0 < N - 2$. Let L_k denote the slope of the line connecting the two points $(i_0, \tilde{\psi}(i_0))$ and $(i_0 + k, \tilde{\psi}(i_0 + k))$. To show that $\tilde{\psi}$ is convex it suffices to show that $L_2 \geq L_1$ and $L_{-1} \geq L_{-2}$. We will just show that $L_2 \geq L_1$ since the proof of the other is similar. Now

$$L_2 - L_1 = \frac{1}{2(i_0+2)} \left(\sum_{j=i_0+3}^{N} x_j \right)^2 + \frac{1}{2i_0} \left(\sum_{j=i_0+1}^{N} x_j \right)^2$$

$$+ \frac{1}{2} x_{i_0+1}^2 - \frac{1}{2} x_{i_0+2}^2 - \frac{1}{i_0+1} \left(\sum_{j=i_0+2}^{N} x_j \right)^2 .$$

Let $d = \sum_{j=i_0+3}^{N} x_j$. Then we have

$$L_2 - L_1 = [i_0(i_0+1)(i_0+2)]^{-1} d^2 + \Big\{ i_0^{-1} x_{i_0+1}$$

$$- (i_0+1)^{-1} x_{i_0+2} + [i_0^{-1} - (i_0+1)^{-1}] x_{i_0+2} \Big\} d$$

$$+ (2i_0)^{-1} x_{i_0+1}^2 + (2i_0)^{-1} x_{i_0+2}^2 + i_0^{-1} x_{i_0+1} x_{i_0+2}$$

$$- (i_0+1)^{-1} x_{i_0+2}^2 + 2^{-1}(x_{i_0+1}^2 + x_{i_0+2}^2)$$

which is greater than zero, so that $\tilde{\psi}$ is convex.

To complete the proof of the theorem we note that for any design $p \in \Phi^1(max)$

$$\sum_{i=1}^{N} \psi(i)p(i) = \sum_{i=1}^{N} \tilde{\psi}(i)p(i) \geq \tilde{\psi}\left(\sum_{i=1}^{N} ip(i)\right) \geq \tilde{\psi}(n)$$

by the convexity of $\tilde{\psi}$ and Jensen's inequality. Let p^* be a design which concentrates all its mass on samples $s \in S_n(max)$. Then $\sum_s n(s)p^*(s) = n$ and it achieves the infimum over Φ^1 for every π^w where $w = 1, \ldots, W$ is the sequence of priors against which $\delta^r_{\mathbf{x}}$ is stepwise Bayes. Since Φ^1 contains designs which put positive mass on every member $s \in S$, $\cup_w \mathcal{Y}^{\mathbf{r}}(\mathbf{b})(\pi^w) = \mathcal{Y}^{\mathbf{r}}(\mathbf{b})$, and we have by the remarks following the proof of Theorem 3.4 that the pair $(p^*, \delta^r_{\mathbf{x}})$ is uniformly admissible. □

The convexity argument used in the above proof is a generalization of an argument due to Joshi (1966) where uniform admissibility results for the sample mean were given.

The designs identified in Theorem 3.5 which yield uniform admissibility are essentially nonrandom or purposeful in nature. For example, when the x_i's are all distinct the theorem yields only one design which observes the n units with the n largest x_i values with probability one. This is not unexpected since a Bayesian approach typically leads to purposeful designs. For more discussion on this point see Basu (1969) and Zacks (1969).

3.5 Nonresponse

The problem of nonresponse is an important one in sample survey and is difficult to handle in the usual frequentist formulation. Various suggestions have been made for imputing values for the missing observations but without much theoretical justification. Theoretically, handling nonresponse is more straightforward from the Bayesian point of view. Practically however it can be more difficult since it requires an additional level of modelling, i.e. modelling how nonresponders are related to the responders.

An interesting approach to this problem is multiple imputation which is described in Rubin (1987). Multiple imputation was developed to handle missing data in public use files where the user has only complete data methods available and has limited information about the reasons for nonresponse. It uses Bayesian ideas to

yield inference procedures with sensible frequentist properties. Rubin distinguishes between ignorable and nonignorable nonresponse. When the nonresponse is nonignorable and followup surveys are not possible, Rubin and others argue, correctly we believe, that any sensible analysis must be based on an assumed model for the nonresponse. Given a sample which contains some nonrespondents he constructs a distribution for the missing observations. A complete data set is then formed, by using this distribution to impute values for all the missing observations. Then the completed data set is analysed using standard procedures just as if the imputed data were the real data obtained from the nonrespondents. This process is repeated several times where in each repetition a new set of imputed values is chosen for the missing observations. This collection of complete data inferences can be combined to form one inference that more properly reflects the uncertainty due to nonresponse than is possible if just one set of imputed values is considered. This is a sophisticated approach which combines frequentist and Bayesian ideas in interesting and unusual ways.

In this section we show how the Polya posterior can be adapted to problems of nonresponse. As in the multiple imputation approach one must be able to construct a model that relates nonrespondents to respondents. Beyond this, however, no prior distribution needs to be specified. Then given the observed members of the sample, one has a predictive distribution for all the unobserved members of the population, both those that are in the sample and those that are not. This predictive distribution can then be used to compute point and interval estimates in the usual Bayesian way. If the assumed model is approximately correct the corresponding estimators should have good frequentist properties.

In Section 3.5.1 we briefly review a Bayesian approach to the problems of nonresponse in a sample survey. In Section 3.5.2 we give Rubin's argument that when the nonresponse is ignorable the Polya posterior is a 'proper' imputation procedure. Almost all of what we say in these two sections is just a gloss on parts of Rubin (1987) and one should check there for the necessary definitions and further details. In Section 3.5.3 we present a modification of the Polya posterior for the problem of nonignorable nonresponse and in Section 3.5.4 apply this method to a population discussed in Greenlees *et al.* (1982). This approach was first given in Meeden and Bryan (1996). Finally, in Section 3.5.5 we briefly compare this method and multiple imputation.

3.5.1 A Bayesian approach

Consider a finite population which contains N units. Associated with unit i let y_i be the value of the characteristic of interest. We consider just the simplest situation where no auxiliary variables are present. Furthermore we are assuming that each unit is either a responder or a nonresponder. We let $t_i = 1$ if the ith unit is a responder and $t_i = 0$ otherwise. For a unit with $t_i = 1$ we are assuming that the value y_i is always observed when the unit is in the sample and if $t_i = 0$ that it is never observed when it is in the sample. Both the y_i's and t_i's are assumed to be unknown to the statistician. So the unknown parameter is the pair of vectors $\mathbf{y}$ and $\mathbf{t} = (t_1, \ldots, t_N)^T$ and a Bayesian statistician must specify a joint prior distribution for the pair $\mathbf{y}$ and $\mathbf{t}$.

In what follows, as always, we let s denote the labels of the $n(s) = n_s$ units which make up our sample. For an $i \in s$ we will observe the corresponding y_i if and only if $t_i = 1$. We let s_r denote the labels of all the units in the sample which are responders and s_{nr} denote the labels of all nonresponders in the sample. Let $n(s_r) = n_r$ denote the number of elements contained in s_r. Then for this problem a typical data point is

$$\begin{aligned} z &= (s, s_r, z_{s_r}) \\ &= (s, s_r, (z_{i_1}, \ldots, z_{i_{n_r}})^T) \end{aligned}$$

where $s_r = (i_1, \ldots, i_{n_r})$ are the labels of the units in the sample of the responders and z_{s_r} gives their values for the characteristic y.

One more bit of notation, in what follows q will always denote a probability density function or probability mass function, conditional or unconditional as the case may be.

Since we will only consider situations where our beliefs about $\mathbf{y}$ and $\mathbf{t}$ are exchangeable the prior we consider will be of the form discussed in Section 1.4. We assume it is given by

$$q(\mathbf{y}, \mathbf{t}) = \int\int \left\{ \prod_{i=1}^{N} q(y_i|t_i, \theta) \right\} g(\lambda) h(\theta)\, d\lambda\, d\theta \qquad (3.17)$$

where

$$q(y_i|t_i, \theta) = \lambda^{t_i} q_R(y_i|\theta) + (1-\lambda)^{1-t_i} q_{NR}(y_i|\theta).$$

Note that the two parameters and four density functions all have straightforward Bayesian interpretations. We see that $q_R(\cdot|\theta)$ is the density function for the characteristic y for a unit which is a re-

sponder conditioned on a value of the mixing parameter θ. $q_{NR}(\cdot|\theta)$ is the analogous density for nonresponders. If we integrate over all the y_i's and θ we see that the resulting distribution for the t_i's is exchangeable. We can think of the parameter λ as representing the proportion of responders in the population and the density $g(\lambda)$ as representing the statistician's prior beliefs about λ. In the same way it is easy to see that the distribution of the y_i for the responders will be exchangeable. The same is true for the y_i for the nonresponders, but these two distributions will be different. How different will depend on the density functions $q_R(y|\theta)$ and $q_{NR}(y|\theta)$. Typically the parameter θ will be a vector and the choice of this parameter along with the two previous conditional densities will be crucial in modelling the relationships between the responders and the nonresponders. The density $h(\theta)$ represents the statistician's prior beliefs about θ and is assumed to be independent of the density g.

Then it is easy to see that for a data point z and a $\mathbf{y}$ and a $\mathbf{t}$ consistent with z we have

$$
\begin{aligned}
&q(\mathbf{y},\mathbf{t}|z) \\
&\quad = \int\int \frac{\prod_{i\in s_r} q_R(y_i|\theta)h(\theta)}{q_R(y_i\,:\,i\in s_r)} \left\{ \prod_{i\in s_{nr}} q_{NR}(y_i|\theta) \right\} \\
&\quad \times \left\{ \prod_{j\notin s} q(y_j|t_j,\theta) \right\} \frac{\lambda^{n_r}(1-\lambda)^{n_s-n_r}g(\lambda)}{q(t_i\,:\,i\in s)}\, d\lambda\, d\theta \\
&\quad = \int\int \left\{ \prod_{i\in s_{nr}} q_{NR}(y_i|\theta) \right\} \\
&\quad \times \left\{ \prod_{j\notin s} q(y_j|t_j,\theta) \right\} g(\lambda|s_r,s_{nr})h(\theta|s_r,z_{s_r})\, d\lambda\, d\theta.
\end{aligned}
$$

If we let s_r' be the complement of s_r then $\mathbf{y}(s_r')$ are just the 'unobserved' units in the population. Again it is easy to see that

$$
\begin{aligned}
&q(\mathbf{y}(s_r'),\mathbf{t}(s')|z) \\
&\quad = \int\int \prod_{i\in s_{nr}} q_{NR}(y_i|\theta) \left\{ \prod_{j\notin s} [\lambda^{t_j} q_R(y_j|\theta) \right. \\
&\quad \left. +(1-\lambda)^{1-t_j} q_{NR}(y_j|\theta)] \right\} g(\lambda|z)h_R(\theta|z)\, d\lambda\, d\theta. \quad (3.18)
\end{aligned}
$$

Note that (3.18) allows us to simulate observations from the posterior distribution of the unobserved given the observed data z in a straightforward manner. For example we would first choose a value λ^* from the density $g(\lambda|z)$. We would then assign values to all the t_i's with $i \in s'$ under the assumption that they are iid Bernoulli(λ^*) random variables. This divides all the units whose labels were not in the original sample into two groups the 'responders' and the 'nonresponders'. Next we would choose a value θ^* from the density $h(\theta|z)$. Then we would assign values to the 'responders' not in the original sample under the assumption that they are iid with common density $q_R(y|\theta^*)$. Finally we would assign values to the nonresponders in the original sample and to the 'nonresponders' not in the original sample under the assumption that they are iid with common density $q_{NR}(y|\theta^*)$.

In the usual Bayesian manner, given a sample, the above process 'imputes' a value to each member of the population whose y value is unknown. This results in a complete simulated version of the population with corresponding population mean under the posterior in (3.18). Suppose we desire an interval estimate of the population mean. If for a given sample we repeat this process a large number of times the resulting collection of computed means is a random sample from the posterior distribution of the population mean. This collection of simulated population means can be used to find an approximate 0.95 Bayesian credible interval in the usual way. The frequency of coverage of such intervals depends not only on the choice of the prior densities g and h but on how well we have modelled the relationship between the responders and nonresponders. In the rest of this section we will consider a stepwise Bayes approach to this problem which uses the Polya posterior. We will consider an example which shows that when our model reflects the relationship between the responders and nonresponders the above interval contains the true mean about 95% of the time.

Finally we note that (3.17) supports another interpretation of nonresponse in the population. Rather than assuming that the population is permanently split into two fixed groups of responders and nonresponders which never change we could assume that each unit has a probability λ of responding when it is included in the sample. In this model under repeated sampling a unit could respond sometimes when it is included in the sample and other times not. In the following we will make use of both interpretations of our model.

3.5.2 The Polya posterior as a proper imputation method

We will now show that when the nonresponse is ignorable that the Polya posterior is a proper imputation method. Definitions of some of the terms used below and further discussion can be found in Rubin (1987).

For simplicity we will assume that the design is simple random sampling without replacement of size n. In Rubin's terminology this means that the sampling mechanism is ignorable. For the rest of this section we will also assume that there is no difference between the responders and the nonresponders, i.e. $q_R(\cdot|\theta) = q_{NR}(\cdot|\theta)$. Under this assumption we can use (3.18) to find the conditional distribution of $\mathbf{y}(s_r')$ given the sample $z = (s, s_r, z_{s_r})$. It is easily seen that

$$\begin{aligned} q(\mathbf{y}(s_r')|z) &= \int \left\{ \prod_{j \notin s_{nr}'} q_R(y_j|\theta) \right\} h_R(\theta|z)\, d\theta \\ &= q(\mathbf{y}(s_r')|s_r, z_{s_r}) \end{aligned}$$

and so the response mechanism is ignorable as well.

Consider now the problem of estimating the population mean $\mu = \mu(\mathbf{y})$ under squared error loss under the assumption that there is no difference between the responders and nonresponders. Then it is easy to show that the mean of the responders in the sample is an admissible estimator of the population mean. This follows since we can modify the argument for the admissibility of the sample mean for the problem without nonresponse given in Section 2.3 to the setup considered here. The resulting stepwise Bayes argument justifies taking $q(\mathbf{y}(s_r')|z)$ to be the Polya posterior for the unseen given the seen. From this it follows that under the Polya posterior

$$E(\mu \mid s_r, z_{s_r}) = \bar{z}_{s_r} \tag{3.19}$$

and

$$\mathrm{Var}(\mu \mid s_r, z_{s_r}) = \frac{N - n_r}{n_r N} \mathrm{Var}(z_{s_r}) \frac{n_r - 1}{n_r + 1} \tag{3.20}$$

where $\bar{z}_{s_r}$ and $\mathrm{Var}(z_{s_r}) = (n_r - 1)^{-1} \sum_{i \in s_r} (z_i - \bar{z}_{s_r})$ are the sample mean and sample variance of the responders in the sample. This means that when we believe there is no difference between the responders and nonresponders and the probability distribution of the response mechanism is exchangeable we can just use the Polya posterior based on the responders in the sample to make

inferences about the entire population in the usual way. So from the theoretical point of view the Polya posterior gives sensible answers when we believe the unseen are like the seen and the unseen includes both nonresponding and unsampled units. For large n_r it will agree with the standard theory which just uses the responders to make inferences and does no imputing.

Recall however that multiple imputation is a solution to a more practical problem. Given a user who only has software that can analyse complete data sets how can one impute the missing values to create several completed data sets that will jointly yield sensible inferences? To understand this approach we consider the special case where we wish to use the sample mean to estimate the population mean.

Let z_s be the y values of all the units that appear in the sample, i.e. the values for both the responders and the nonresponders. Even though we will never completely know z_s when there is nonresponse it plays an important role in the motivation of multiple imputation. We let $\delta^{sm}(z_s) = \bar{z}_s = \sum_{i\in s} z_i/n$. If z_s were known we could then use $\bar{z}_s$ to estimate μ with estimated variance $\mathrm{Var}(z_s)(1 - n/N)/n$ where $\mathrm{Var}(z_s) = (n-1)^{-1}\sum_{i\in s}(z_i - \bar{z}_s)$ is the sample variance. Under the usual asymptotic theory confidence intervals based on this estimate and its estimated variance are known to have frequency of coverage approximately equal to the stated coverage probability. Or in Rubin's terminology the standard procedure has 'randomization validity'.

To find an interval estimate of μ using the Polya posterior based on the responders one can simulate copies of the entire population, just as we did earlier, or when n_r is large use the normal approximation to the Polya posterior with the variance given in (3.20). There is another way to find this interval estimate approximately which is closely related to the first method. This second is based on an alternative expression for $\mathrm{Var}(\mu|s_r, z_{sr})$ which follows from the well-known formula which gives a variance as the sum of the variance of a conditional expectation with the expectation of a conditional variance. Using this fact we have

$$\begin{aligned}
&\mathrm{Var}(\mu \mid s_r, z_{s_r}) \\
&\quad = \mathrm{Var}\{E(\mu \mid s, z_s) \mid s_r, z_{s_r})\} + E\{\mathrm{Var}(\mu \mid s, z_s) \mid s_r, z_{s_r})\} \\
&\quad = \mathrm{Var}(\bar{z}_s \mid s_r, z_{s_r}) + \frac{n-1}{n+1}\frac{N-n}{nN} E\{\mathrm{Var}(z_s) \mid s_r, z_{s_r}\}
\end{aligned} \tag{3.21}$$

where the second line uses (3.19) and (3.20) but with s_r replaced by s.

Now let us suppose for a moment that we wish to estimate $\bar{z}_s$ using z_{s_r}. This is just the standard finite population problem of estimating the population mean where the population is now z_s and we have observed the sample (s_r, z_{s_r}). Hence we can use the Polya posterior based on the responders in a straightforward manner to solve this problem. In fact the form of (3.19) and (3.20) for this problem are

$$E(\bar{z}_s \mid s_r, z_{s_r}) = \bar{z}_{s_r} \tag{3.22}$$

and

$$\mathrm{Var}(\bar{z}_s \mid s_r, z_{s_r}) = \frac{n - n_r}{n_r n} \mathrm{Var}(z_{s_r}) \frac{n_r - 1}{n_r + 1}. \tag{3.23}$$

The Polya posterior can be used to impute values for the non-responders in the sample in the usual Bayesian way to construct a 'complete' sample, say $z^*_{s,1}$. We could then repeat this process an additional $m - 1$ times to form $z^*_{s,1}, \ldots, z^*_{s,m}$ completed samples. Then the average of the means of these m completed samples should be approximately $\bar{z}_{s_r}$ and be a good estimate of $\bar{z}_s$ and hence a good estimate of μ. In the same way we can use the m simulated completed samples to estimate the two terms in the second line of (3.21) and hence estimate $\mathrm{Var}(\mu | s_r, z_{s_r})$. That is rather than use the Polya posterior, based on the responders, to simulate completed copies of the entire population one can just use it to simulate m completed samples and get approximate estimates of the posterior mean and posterior variance of μ given (s_r, z_{s_r}). Hence multiple imputation, i.e. the creation of these m imputed or completed samples, can be thought of as an approximation to a Bayes or stepwise Bayes analysis.

The above development gives a way to think about multiple imputation as an approximation to a standard Bayesian posterior analysis where one uses (s_r, z_{s_r}) to estimate $\bar{z}_s$ when (s, s_r, z_{s_r}) is considered fixed. Rubin's approach is slightly different. He begins with a standard frequentist procedure, say $\delta^{sm}(z_s)$, with its estimated variance $\mathrm{Var}(z_s)(1 - n/N)/n$ which has randomization validity. He then wants to find a Bayesian model such that if one uses the model to constuct completed samples the resulting multiple imputation approximation to $\mathrm{Var}(\mu | s_r, z_{s_r})$ of the Bayesian model will also be a good approximation to the estimated variance

of the frequentist procedure under consideration. In this case the resulting inferential procedure will still have good frequentist properties, i.e. will also have randomization validity. In order for this to occur the Bayesian model for the unseen give the seen and the response mechanism must satisfy certain conditions. When these assumptions hold the imputation procedure is said to be proper. So for him the Bayesian model is essentially a handmaiden that is used to modify or correct a reasonable frequentist procedure when nonresponse is present. In what follows we always assume that the frequentist procedure of interest is δ^{sm}, the sample mean.

The first set of conditions is on the posterior distribution of the Bayesian model and requires that for a fixed complete sample, z_s, the posterior mean and variance are asymptotically equal to $\bar{z}_s$ and $\mathrm{Var}(z_s)(1 - n/N)/n$ respectively. But we have already seen that this is true for the Polya posterior in Section 2.6. The second set of conditions is on the response mechanism and requires some additional notation. From the frequentist perspective if we imagine drawing repeated samples for this problem then not only would the samples change from trial to trial but for a fixed sample s the sample of responders, s_r, could vary within the subsequence of trials where s is observed. How the s_r vary will depend on the underlying response mechanism. Hence for a fixed s we can think of the response mechanism as defining a sampling design over the subsets of s. We will assume that this design is exchangeable in the sense that for any two choices of s_r that are of the same size will also have the same probability of being selected. Therefore conditional on the event that $n_r = n_1$ this design is like simple random sampling without replacement of size n_1 from s. We denote this induced sampling design over subsets of s by $p_{s,R}$. The second set of conditions is for the problem of estimating the mean of the complete sample z_s, where s and $\mathbf{y}$ are considered to be fixed, and involve taking expectations with respect to the design $p_{s,R}$. It is important to keep in mind that for any such expectation any function of z_s is also fixed.

Recall that for this problem $\bar{z}_{s_r}$ was our estimate for $\bar{z}_s$ under the Polya posterior. Note that $\bar{z}_{s_r}$ is a function of s_r and under this design is a random variable. The first requirement for this second set of conditions is that

$$E_{p_{s,R}}(\bar{z}_{s_r} \mid s, \mathbf{y}) = \sum_{s_r \subseteq s} \bar{z}_{s_r} p_{s,R}(s_r) = \bar{z}_s.$$

This is true because $p_{s,R}$ is exchangeable and the sample mean is unbiased for estimating the population mean for simple random sampling of a fixed size. The result follows by conditioning on the size of s_r.

The other two conditions involve the last two terms of (3.21). Now it is easy to see (Ghosh and Meeden, 1983) that under the Polya posterior

$$\frac{N-n}{nN}E\{\mathrm{Var}(\bar{z}_s) \mid s_r, z_{s_r}\} = \frac{N-n}{nN}\frac{n_r-1}{n_r+1}\frac{n+1}{n-1}\mathrm{Var}(z_{s_r}).$$

The second condition is that $p_{s,R}$ expectation of the above be approximately equal to the estimated variance of $\delta^{sm}(\bar{z}_s)$, i.e.

$$\frac{N-n}{nN}\frac{n+1}{n-1}\sum_{s_r\subseteq s}\frac{n_r-1}{n_r+1}\mathrm{Var}(z_{s_r})p_{s,R}(s_r) \doteq \frac{N-n}{nN}\mathrm{Var}(z_s).$$

By the same argument as given in the above for the mean we have that under $p_{s,R}$ $\mathrm{Var}(z_{s_r})$ is an unbiased estimator of $\mathrm{Var}(z_s)$ since $p_{s,R}$ is exchangeable. Moreover if n_r and n are assumed to be the large (which is all that Rubin requires) then there is approximate equality in the above expression.

As we have just seen, $\bar{z}_{s_r}$ is an unbiased estimator of $\bar{z}_s$ under the design $p_{s,R}$. Let $\mathrm{Var}_{p_{s,R}}(\bar{z}_{s_r})$ denote its variance. The third condition is that the $p_{s,R}$ expectation of $\mathrm{Var}(\bar{z}_s|s_r, z_{s_r})$ is approximately equal to $\mathrm{Var}_{p_{s,R}}(\bar{z}_{s_r})$. By (3.23) this is equivalent to

$$\sum_{s_r\subseteq s}\frac{n-n_r}{n_r n}\frac{n_r-1}{n_r+1}\mathrm{Var}(z_{s_r})p_{s,R}(s_r) \doteq \mathrm{Var}_{p_{s,R}}(\bar{z}_{s_r}).$$

Just as for the previous condition we can see that we will have approximate equality here under the assumption that n_r is large and n is large compared to n_r by conditioning on the value of n_r since in this case both sides just involve expectations under simple random sampling without replacement.

A multiple imputation scheme is said to be 'proper' if it satisfies the above two sets of conditions. The second set of these conditions concerns the problem of estimating z_s when s and $\mathbf{y}$ are fixed and $p_{s,R}$ is the design. For the Polya posterior the conditions are that $\bar{z}_{s_r}$ is unbiased for estimating $\bar{z}_s$ and $\mathrm{Var}(z_{s_r})$ is approximately unbiased for estimating $\mathrm{Var}(z_s)$ and $\mathrm{Var}_{p_{s,R}}(\bar{z}_{s_r})$. The validity of these conditions depends both on the Polya posterior and the underlying response mechanism. To see that some assumptions are needed on the response mechanism consider the following exam-

ple. Suppose that the response mechanism is such that, for any s, s_r is just those labels which are less than $N/2$. In this case we could only observe units whose labels put them in the first half of the population. Clearly such a response mechanism is not exchangeable and will have poor frequentist properties for many $\mathbf{y}$ in a typical parameter space.

Rubin proves, using standard asymptotic theory, that if the multiple imputation is proper and the complete data inference is randomization valid then the multiple imputation procedure is randomization valid as well. In such cases by considering imputed, completed samples an unsophisticated data analyst can use standard software packages to produce an inferential procedure which reflects the additional uncertainty caused by the nonresponse.

The example we have considered here is the simplest one possible, i.e. assuming that the responders are like the nonresponders and an exchangeable response variable. Rubin considers more complicated situations where modelling the nonresponse becomes more crucial. The Polya posterior can be extended to more complicated situations as well. If covariates are present one possibility is to use the covariates to identify subclasses of units such that within each subclass you believe the responders are like the nonresponders. Then within each subclass you could use the Polya posterior to carry out either a Bayesian analysis or multiple imputation. Another possibility is to develop a flexible modelling approach that would modify the Polya posterior for responders to yield a predictive distribution for the nonresponders for a variety of beliefs about how the two groups are related. Then in more realistic problems the two approaches could be combined. In the next section we will consider a modelling approach that gives the nonresponders their own predictive distribution.

3.5.3 Adapting the Polya posterior

We now specialize the model of Section 3.5.1 to yield an approach, based on the Polya posterior, to the problem of nonresponse when one's beliefs about the responders and nonresponders are not exchangeable. We begin by assuming that the characteristic y can take on only finitely many values, say k, where $k \geq 2$. This is not necessary for most of what follows but is done for ease of exposition. We let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ where each θ_i is nonnegative and their sum is one. Next we assume that $q_R(y|\boldsymbol{\theta})$ is the multinomial

probability function with parameters 1 and $\boldsymbol{\theta}$. That is, given $\boldsymbol{\theta}$ and all the t_i's the conditional distribution of the responders is like a random sample from a multinomial(1, $\boldsymbol{\theta}$) distribution.

In the next step we must model the relationship between the responders and the nonresponders. Let C denote a square matrix of order k whose every element is nonnegative and where every column sums to one. By the properties of C if $\boldsymbol{\nu} = C\boldsymbol{\theta}$ then $\boldsymbol{\nu}$ is also a probability vector, that is each of its components is nonnegative and they sum to one. We will assume that $q_{NR}(y|\boldsymbol{\theta})$ is the multinomial probability function with parameters 1 and $C\boldsymbol{\theta}$. That is, given $\boldsymbol{\theta}$ and all the t_i's the conditional distribution of the nonresponders is like a random sample from a multinomial(1, $C\boldsymbol{\theta}$) distribution. Hence, in this approach, it is the matrix C which models the relationship between the responders and nonresponders and must be chosen to reflect the assumed relationship between these two groups.

To help understand how a statistician would use prior information to choose C we consider the following scenario. Suppose the population was completely known to us. We could then split it into two groups, the responders and the nonresponders. Let $\hat{\theta}_i$ be the proportion of responders that fall into category i and $\hat{\nu}_i$ be the proportion of nonresponders that also fall into this category. In general there will be many matrices C for which $\boldsymbol{\nu} = C\boldsymbol{\theta}$. This just reflects the fact that there are many different models for the relationship between the responders and nonresponders which can lead to a specific population. If in fact we believed that a certain model was operating in the population of interest, then we could study it to determine how $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ are related. We would then choose a matrix C which reflects these relationships. Although such an approach is always possible in principle we believe that in practice one will usually not have enough prior information to carry it out successfully. However, one can still choose an appropriate matrix C when only the more typical and less exact prior information is at hand.

For example suppose we believe, for whatever reasons, that the population of nonresponders is stochastically larger than the population of responders. Furthermore, suppose we believe that if 20% of the probability below the median in the population of responders was shifted in an appropriate manner to above the median the resulting population would be very similar to the population of nonresponders. Assuming we have a reasonable prior guess for

the median of the population of responders one can easily choose a matrix C which reflects these prior beliefs. Note that in this crude approach we are not making an attempt to model the actual mechanism of nonresponse. All that is needed is some idea of how the populations of responders and nonresponders compare. This seems to be the minimum amount of prior information required for any sensible approach to the problem of nonresponse. Without such information it is unlikely that much can be done.

It remains to choose the prior densities g and h. This will be done in a noninformative manner. As we have seen when using a multinomial distribution the convenient family of priors is the Dirichlet family. If we assume that the prior for $\boldsymbol{\theta}$ is Dirichlet with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$ then given the sample the posterior distribution of $\boldsymbol{\theta}$ is Dirichlet with a parameter vector whose ith coordinate is $\alpha_i + c_z(i, s)$ where $c_z(i, s)$ is the number of responders in the sample which fall into category i, i.e. the number of times the ith category appears in the set z_{s_r}. If it were the case that each $c_z(i, s) \geq 1$ then we could take the vector $\boldsymbol{\alpha}$ to be zero and still have a sensible posterior distribution. In fact the stepwise Bayes argument of Section 2.3.2, which proved the admissibility of the sample mean, carries over in a straightforward way to the model for nonresponse given above and will be omitted. In particular it leads to the pseudo posterior for $\boldsymbol{\theta}$ which comes from assuming that all the α_i's are zero. In fact one can let the matrix C depend on the categories observed in the sample and the argument still works. This fact will be used in the next section where an example is considered. Hence, other than choosing the matrix or matrices C one need not specify a prior for $\boldsymbol{\theta}$. To summarize, given the sample z, one may take as the posterior for $\boldsymbol{\theta}$ a Dirichlet distribution with the parameter vector whose ith coordinate is $c_z(i, s)$. Hence one need not ever specify the density h.

In much the same way one need not specify the density g. The natural choice is a beta density. Now for any sample with both responders and nonresponders present we can take the two parameters of the beta density to be zero and the resulting pseudo posterior is a beta density with parameters $(n_r, n - n_r)$. This noninformative choice of g can also be given a stepwise Bayes justification. In conclusion we see that we only need to specify approximately the relationship between the responders and nonresponders to be able to implement a stepwise Bayes analysis for this problem. In the next section we will see how this can be done in practice.

3.5.4 An example

In this section we will consider an interesting data set which involved nonresponse and where the true value of the nonresponders could be determined. This data set was presented in Greenlees *et al.*, (1982). The Current Population Survey is a large monthly survey conducted by the Census Bureau to provide data on yearly income and other characteristics. Using information from the IRS and social security numbers the true income for every member in the sample for March 1973, including nonrespondents, was found. For various reasons certain individuals were removed from the sample until a final population of 5515 units was constructed. For these 5515 individuals their true yearly income was known along with whether or not they had responded to the original survey. Of these 410 had refused to respond to the original survey and 151 had left the question unanswered for other reasons.

They ignored these last 151 individuals and hence considered a population with 5364 individuals. They hypothesized that the probability of response depended on income and other auxiliary variables through a logistic model. Their analysis showed that the probability of response did depend on income and indicated that individuals with higher income have smaller probabilities of response. They then consider models, which used the auxiliary variables, to impute the missing values. They showed that their methods did a better job imputing the missing values than methods which assumed that the nonresponse was ignorable.

Following them we will consider the population with 5364 individuals where 410 are nonrespondents. Although it is not necessary we simplified matters somewhat by rounding to construct 26 possible values for yearly income. All individuals whose income was \$1999 or less were assigned the value of 1. All individuals whose income exceeded \$1999 but did not exceed \$3999 were assigned the value of 3. We continued in this way for all incomes which did not surpass \$50 000. Since in the original data all incomes which exceeded \$50 000 were censored we assigned them the value 51. This gives us the 26 possible values. To compare the populations of responders and nonresponders study the first two plots of Figure 3.10 which gives their bar plots. Inspection of the two plots shows that indeed the distribution of nonresponders puts more mass on the larger income values than does the distribution of the responders. In fact the mean of the responders is 13.52 while the mean of the

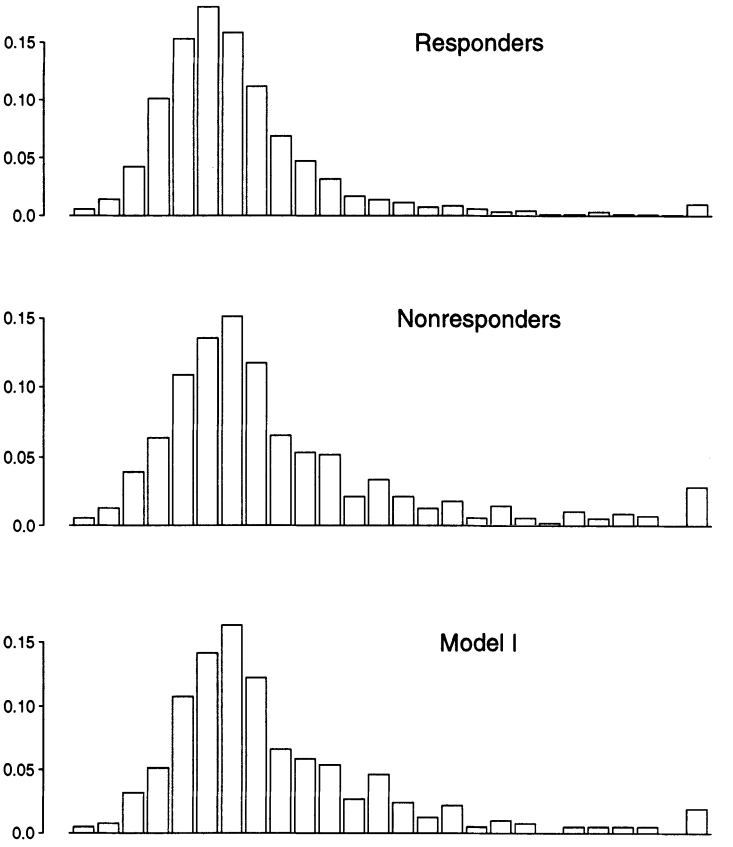

Figure 3.10 *Bar plots of the population of responders, the population of nonresponders and the population of nonresponders under model I.*

nonresponders is 16.47. Moreover the sum over the categories of the absolute value of the difference between the relative frequencies is 0.31.

Now in an actual problem such complete and explicit information about the nonrespondents will not be available to the statistician. But as a first check we wanted to select a matrix C for which our model for the nonrespondents seemed to be a good approximation for the actual population of nonrespondents. If our approach does

not work in such a case then it is hard to imagine it could be of any use in more realistic problems. One such matrix C can be found as follows. We begin by assuming the probability of falling in category 1 is the same for both populations. For the remaining 25 categories we assume that we can get from the distribution of responders to the distribution of nonresponders by removing 30% of the probability from the categories 2 to 6 and spreading 80% of this probability uniformly over categories 7 to 16 and the remaining 20% uniformly over categories 17 to 26. More formally the relationship between $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ is given by

$$\nu_i = \begin{cases} \theta_1 & \text{if } i = 1 \\ 0.7\theta_i & \text{if } i = 2, \ldots, 6 \\ \theta_i + 0.024(\theta_2 + \cdots + \theta_6) & \text{if } i = 7, \ldots, 16 \\ \theta_i + 0.006(\theta_2 + \cdots + \theta_6) & \text{if } i = 17, \ldots, 26 \end{cases}$$

The plot in Figure 3.10 gives the histogram for the model of the nonresponders defined by the transformation given above. Clearly the histogram based on the model is more like the histogram for the nonresponders than the histogram for the responders. The mean of the distribution for the model is 16.08 and the sum of the absolute differences between the relative frequencies in the histograms for the nonresponders and the model is 0.13.

To see how the relationship assumed above would improve on the interval estimation of the population mean we took 1000 simple random samples without replacement of size 500 from the entire population of 5364 individuals. We calculated the usual 95% interval for each sample just ignoring the missing observations. We also found the 0.95 credible interval, based on the Polya posterior, described here using the above matrix C. For each sample we simulated 1000 completed populations to form our interval. We see from Table 3.8 that for this model, model I, and for these 1000 samples the average length of the usual interval was 1.30 and it contained the true mean 87% of the time, while the average length of the modified Polya posterior interval was 1.23 and it contained the true mean 93% of the time.

It would be easy, however, to make other choices for the matrix C which would result in the assumed model for the nonresponders being much more similar to the actual population of nonresponders than what occurred in this case. Such choices would require much more prior information than would usually be available. To use the relationship described above one only needs to know that the re-

Table 3.8 *Comparison of the usual 95% interval and some 0.95 modified Polya posterior intervals for 1000 random samples of size 500.*

Model	Type	Average length	Relative frequency of coverage
I	Usual	1.30	0.870
	Polya	1.23	0.930
II	Usual	1.31	0.895
	Polya	1.23	0.957
III	Usual	1.30	0.887
	Polya	1.24	0.913
IV	Usual	1.30	0.878
	Polya	1.25	0.899

sponders tend to be smaller than the nonresponders and have some idea of how different the two populations are. To see how robust our method is we considered several other choices of C and ran similar simulation studies. We tried to select less complicated C's, and ones that realistically could be used in practice. For example, in model II we chose the C which takes 25% of the probability from each of the first six categories and spreads it uniformly over the remaining categories. In model III we took 21% of the probability from the first six categories and spread it uniformly over categories 8 to 13. All three of these choices for C result in a model for the nonresponders which is quite different from the responders. However, the last two did not require as much detailed information about the relationship between the responders and the nonresponders as the first model did. For a more conservative approach to the problem, in model IV we took 20% of the probability from each of the first 25 categories and moved it to the next higher category.

For each of these last three models the performances of the usual 95% interval and the 0.95 interval of the modified Polya posterior were compared for 1000 random samples of size 500. We see from Table 3.8 that the performance of all four of the models is better than that of the usual interval. As was to be expected the amount

of improvement is smallest for model IV since it was the most conservative of the four models. These results suggest that this approach is rather robust against the choice of C and any sensible choice of C should yield some improvement. For example, in this problem as long as we move some probability to the right and not to the left we should improve upon the usual interval.

As we have noted, rounding the units so that only a relatively few values appear in the sample and the assumption that these values are known a priori to the statistician are not necessary to implement the analysis described here. For example, for the population of responders the probability of falling into the first six categories is 0.4963. Hence an essentially equivalent analysis can be carried out in the following manner when neither the number of categories nor their values is known. However, we assume that a prior guess for the median of the responders, say m, is known. Given a sample we find the categories that appear in the sample. We then find the matrix which moves 30% of the probability below the value m to above m, and places 80% of this probability uniformly on the smaller half of the categories above the value m and the remaining 20% of this probability on the larger half of the categories above the value m. The matrix will depend on the categories that appear in the sample but not on their relative frequencies and so the stepwise Bayes argument can proceed in the usual fashion.

3.5.5 Discussion

We believe that the example presented in the previous section demonstrates that a stepwise Bayes approach to nonignorable nonresponse given here can be useful in practice. To implement it one needs no more prior information than any other sensible approach to this problem requires. It is also quite flexible. It incorporates easily various types of prior information about the relationship between the responders and nonresponders. It generalizes in a straightforward way to situations where one can identify various classes of responders and nonresponders when information from auxiliary variables is available. Its flexibility makes it easy to study the sensitivity of the analysis to the model for nonresponse. In addition it has all the advantages of a fully Bayesian analysis. One obtains a posterior distribution for all the unobserved units whether they were in the sample or not. This allows one to impute

values to all the unknown units and compute point and interval estimates in the usual Bayesian manner. If the model chosen approximately reflects the relationship between the responders and the nonresponders the resulting interval estimates will have good frequency properties as well.

The problem of nonresponse in sample surveys seems to be one where standard frequentist methods are quite difficult to apply. Chiu and Sedransk (1986) consider a Bayesian procedure for imputing missing values in the sample. They assume that the characteristic y_i can take on finitely many values, say $\mathbf{b} = (b_1, \ldots, b_k)^T$ and for a given unit i with $y_i = b_j$, the probability of nonresponse depends on b_j. Their approach is quite flexible and permits the inclusion of various types of prior information. However, they just consider the problem of imputing the missing observations in the sample and never consider a posterior distribution for all the unobserved units.

The approach given here is very similar in spirit to that of Rubin (1977). In both cases a predictive distribution for the unobserved given the sample is found. However, the technology of producing the predictive distribution is quite different. Recently Rubin has been more interested in multiple imputation. As we have seen this approach to the problem of nonresponse is an interesting mix of frequentist and Bayesian ideas. It has always seemed a bit surprising to us that Rubin just finds a posterior distribution for the nonresponders. We are not objecting as a frequentist might to the notion of such a posterior. This, however, seems to us to be the most difficult part of the problem and if one can find this posterior then one should be able to find a posterior for the responders as well since we, presumably, know more about them. In fact for us this second posterior is indeed easy to find and we essentially use it to find a posterior for the nonresponders. His ultimate justification for multiple imputation is that it leads to procedures with improved frequentist properties for problems where complete samples are necessary to satisfy the user. The method which we have suggested here is similar in spirit to Rubin's multiple imputation approach. Indeed one could use our posterior for the nonresponders to repeatedly impute just the values that are missing in the sample. We have argued though for a more thoroughgoing Bayesian approach than multiple imputation. Although the two approaches could often give very similar answers we believe that the approach

suggested here has a more straightforward Bayesian interpretation and can lead to procedures with good frequentist properties as well.

3.6 Some nonparametric problems

Although finite population sampling is clearly a distinct area of statistical inference, it sometimes seems more isolated from the mainstream than it should be. For example, it seems to us that many problems which are routinely considered to be nonparametric could just as easily be formulated as finite population sampling problems. The choice seems to depend on what kind of prior information is assumed to be available. In this section we will see that there is a close relationship between certain kinds of questions in finite population sampling and nonparametric problems. In Section 3.6.1 we will show that the stepwise Bayes argument which gives the admissibility of procedures based on the Polya posterior can easily be adapted to prove the admissibility of standard nonparametric estimators. This follows since in both cases the admissibility questions can be reduced to proving admissibility in a multinomial problem. In Section 3.6.2 the relationship between the Polya posterior and the Dirichlet process priors (Ferguson, 1973) and the Bayesian bootstrap (Rubin, 1981) will be discussed. In Section 3.6.3 we will see results for the problem of nonparametrically estimating a median that are very similar to those for estimating a median in finite population sampling given in Section 2.8.2.

3.6.1 Proving admissibility

Let $Z_1, \ldots, Z_n$ be a random sample from an unknown distribution function F, which is assumed to belong to $\mathcal{F}$, some large nonparametric family of distributions on the set of real numbers. We wish to estimate

$$\gamma(F) = \int \phi(t)\, dF(t) \tag{3.24}$$

where ϕ is some specific function, with squared error loss. We will assume that $\mathcal{F}$ is the class of all distribution functions F for which the integral in (3.24) is finite. We will consider the question of finding admissible estimators for this problem. It turns out that sometimes this can be accomplished by finding admissible estimators for the following simpler problem.

Let $\mathbf{b} = (b_1, \ldots, b_k)^T$ be a vector of k distinct real numbers. Let

$\mathcal{F}(\mathbf{b})$ denote all those distribution functions which concentrate all their mass on the points $b_1, \ldots, b_k$. We now consider the problem of estimating $\gamma(F)$ when F is assumed to belong $\mathcal{F}(\mathbf{b})$. In this case $Z_1, \ldots, Z_n$ is a random sample from a multinomial(1; $\boldsymbol{\theta}$) where $\theta_i = P(Z_j = b_i)$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n$. Note that $\mathcal{F}(\mathbf{b})$ is equivalent to the $(k-1)$-dimensional simplex

$$\Lambda = \left\{ \boldsymbol{\theta} : \theta_i \geq 0 \text{ for } i = 1, \ldots, k \text{ and } \sum_{i=1}^{k} \theta_i = 1 \right\}. \tag{3.25}$$

If $z = (z_1, \ldots, z_n)^T$ denotes the vector of a possible set of outcomes for a random sample of size n, then we let

$$c_z(i, n) = \text{number of } z_j\text{'s in } z \text{ which equal } b_i$$

for $i = 1, \ldots, k$.

Recall that in Section 2.3.2, to prove the admissibility of the sample mean for estimating the population mean in finite population sampling when the parameter space is $\mathcal{R}^N$, it was enough to prove its admissibility when the parameter space is $\mathcal{Y}(\mathbf{b})$ for an arbitrary vector $\mathbf{b}$. We shall soon see that a similar relation holds for the parameter spaces $\mathcal{F}$ and $\mathcal{F}(\mathbf{b})$ for the nonparametric problem considered here. Moreover in each case the admissibility result for the smaller parameter space follows from the same stepwise Bayes argument. In other words, once an admissibility result can be demonstrated for a multinomial problem it can often easily be extended to yield admissibility results for both finite population sampling and nonparametric problems.

Let π be a prior distribution for the multinomial problem with parameter space $\Lambda = \mathcal{F}(\mathbf{b})$ defined in (3.25). Given a sample $z = (z_1, \ldots, z_n)^T$ the Bayes estimate of θ_i against π is

$$\begin{aligned} E_\pi(\theta_j \mid z) &= \int \cdots \int \theta_j \prod_{i=1}^{k} \theta_i^{c_z(i,n)} \, d\pi(\theta_1, \ldots, \theta_k) \\ &\div \int \cdots \int \prod_{i=1}^{k} \theta_i^{c_z(i,n)} \, d\pi(\theta_1, \ldots, \theta_k) \\ &= P_\pi(b_j \mid z) \end{aligned}$$

which is the π posterior probability that an additional observation

takes of the value b_j. From this it follows that

$$E_\pi(\gamma \mid z) = \sum_{i=1}^{k} \phi(b_i) P_\pi(b_i \mid z) \tag{3.26}$$

is the Bayes estimate of γ against π given z.

We now turn to the analogous finite population sampling estimation problem. We will be considering two cases, when the parameter space is either $\mathcal{R}^N$ or $\mathcal{Y}(\mathbf{b})$. In both of the cases if $\mathbf{y}$ is a typical parameter point we denote the distribution function which assigns mass $1/N$ to each point y_i of $\mathbf{y}$ by $F_\mathbf{y}$. We consider the problem of estimating

$$\gamma(\mathbf{y}) = \gamma(F_\mathbf{y}) = \int \phi(t)\, dF_\mathbf{y}(t) = \sum_{i=1}^{N} \phi(y_i)/N$$

under squared error loss. Now any prior π on Λ induces a prior distribution π^* over $\mathcal{Y}(\mathbf{b})$ by the relationship

$$\pi^*(\mathbf{y}) = \int \cdots \int \prod_{i=1}^{k} \theta_i^{c_y(i)}\, d\pi(\theta_1, \ldots, \theta_k)$$

for $\mathbf{y} \in \mathcal{Y}(\mathbf{b})$ and where $c_y(i)$ and $c_z(i, s)$, used below, were defined in the proof of Theorem 2.3. If $z = (s, z_s)$ is the observed data under some design p with $n(s) = n$ then the Bayes estimate for γ is

$$E_{\pi^*}(\gamma \mid z) = \frac{n}{N} \sum_{i=1}^{k} \phi(b_i) c_z(i, s)/n + \frac{N-n}{N} E_\pi(\gamma \mid z_s). \tag{3.27}$$

Note that when we write $E_\pi(\gamma|z_s)$ we are identifying z_s, the observed values in our sample of size n, as a vector of observations of length n for the multinomial problem. Hence we are letting z stand for two different things. In the multinomial or nonparametric problem it is just the vector of observed values while in the finite population sampling problem it is the set of labels of the units in the sample along with their observed values.

Note that there is an interesting relationship between the estimators in (3.26) and (3.27). Suppose that we are given a set of n observations whose values belong to the set $\mathbf{b}$. If we assume that the observations arose from the multinomial model with prior π, then the estimator in (3.26) is the proper one. If, on the other hand, we assume that the observations were generated by the fi-

nite population model with prior π^*, then the estimator in (3.27) is the proper one. This suggests that the term $\sum_{i=1}^{k} \phi(b_i)c_z(i,s)/n$ in (3.27) can be interpreted as the finite population correction factor to the estimator given in (3.26). If we let F_{z_s} denote the distribution function which puts mass $1/n$ on each member of z_s and

$$\delta_{\pi^*}(z) = E_{\pi^*}(\gamma \mid z) \text{ and } \delta_{\pi}(z) = E_{\pi}(\gamma \mid z)$$

then (3.27) can be rewritten as

$$\delta_{\pi^*}(z) = \frac{n}{N} \int \phi(t)\, dF_{z_s} + \frac{N-n}{N} \delta_{\pi}(z). \tag{3.28}$$

If π is such that δ_π is the unique Bayes estimator against π for the multinomial problem with parameter space $\Lambda = \mathcal{F}(\mathbf{b})$, then δ_π is admissible. In addition δ_{π^*} is the unique Bayes estimator against π^* for the finite population sampling problem with parameter space $\mathcal{Y}(\mathbf{b})$ and hence δ_{π^*} is admissible. More generally, if δ is the unique stepwise Bayes estimator against a sequence of priors for the multinomial problem with parameter space $\lambda = \mathcal{F}(\mathbf{b})$, then δ is admissible. In addition, the estimator

$$\delta^*(z) = \frac{n}{N} \int \phi(t)\, dF_{z_s} + \frac{N-n}{N} \delta(z) \tag{3.29}$$

is a unique stepwise Bayes estimator for the finite population sampling problem with parameter space $\mathcal{Y}(\mathbf{b})$ and hence admissible.

In the proof of Theorem 2.4 we saw that for the finite population sampling problem it was an easy matter to demonstrate the admissibility of an estimator when the parameter space is $\mathcal{R}^N$ if it is known to be admissible when the parameter space is $\mathcal{Y}(\mathbf{b})$ for each possible vector $\mathbf{b}$. In exactly the same way the admissibility of an estimator for the multinomial problem with parameter space $\Lambda = \mathcal{F}(\mathbf{b})$ for each possible vector $\mathbf{b}$ yields the admissibility of the estimator for the nonparametric problem with parameter space $\mathcal{F}$. This was first noted by Cohen and Kuo (1985), who demonstrated that the empirical distribution function is an admissible estimator of F itself for a certain nonparametric problem. Brown (1988) used similar arguments to study the admissibility of various nonparametric estimators. In the next section we will consider some other examples.

3.6.2 Some examples

In this section we will give a few examples of admissible nonparametric procedures and explore the relationship between methods based on the Polya posterior and the Dirichlet process priors (Ferguson, 1973) and the Bayesian bootstrap (Rubin, 1981). In the examples that follow it will always be the case that for the sample s we have $n(s) = n$.

Example 1. The sequence of priors which proved the admissibility of the sample mean in Theorem 2.3 also can be used to prove the admissibility of $\sum_{i=1}^{n} \phi(z_i)/n = \overline{\phi(z)}$ for estimating $\gamma(F)$ in the nonparametric problem and $\sum_{i\in s} \phi(z_i)/n$ for estimating $\sum_{i=1}^{N} \phi(y_i)/N$ in the finite population sampling problem. The same sequence of priors was used in Brown (1981) (see also Alam (1979)) in the special case where $\phi(t) = t$.

Example 2. Let ϕ be an arbitrary nonconstant Borel measurable function with $m \in (\inf_t \phi(t), \sup_t \phi(t))$ and $w > 0$ given. Then using the sequence of priors used in the proof of Theorem 3.1 it follows that the estimator

$$\frac{wm}{w+n} + \frac{n\overline{\phi(z)}}{w+n} \tag{3.30}$$

is admissible for estimating $\gamma(F)$ in the nonparametric problem. In the special case $\phi(t) = t$ the finite population sampling analogue of (3.30) is the estimator $\delta_{m,w}$ defined in (3.1). Also in this special case the estimator defined in (3.30) is the estimator given in equation (7) of Ferguson (1973) if one identifies w with the parameter $\alpha(\mathcal{R})$ of the Dirichlet process.

Example 3. The arguments of the preceding section can be generalized to obtain admissible estimators of parameters $\gamma(F)$ of types other than those which satisfy (3.24). For example, let $\gamma(F)$ be an estimable parameter with kernel ϕ and degree $v > 1$, i.e.

$$\gamma(F) = \int \cdots \int \phi(z_1, \ldots, z_v)\, dF(z_1) \cdots dF(z_v)$$

where, without loss of generality, it can be assumed that ϕ is symmetric in its arguments. One special case of interest for $v = 2$ is $\phi(z_1, z_2) = (z_1 - z_2)^2/2$. Using the sequence of priors of Example 1, it follows that $(n+1)^{-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2$ is an admissible estimator of the variance of F.

It is interesting to note that if one assumes the population to be normal then this estimator is the best invariant estimator of

the variance of F. However, as shown by Stein (1964), such an estimator is not admissible under squared error loss.

If the sequence of priors of Example 2 is used then one obtains, as an admissible estimator of the variance of F, the estimator given in equation (15) of Ferguson (1973). For the finite population sampling versions of these two see Ghosh and Meeden (1983) and Vardeman and Meeden (1983).

For some additional examples see Meeden *et al.*(1985) In Meeden *et al.*(1989), essentially this same argument, using the sequence of priors of Example 1, proved the admissibility of the Kaplan–Meier estimator. The argument was somewhat more complicated however because of the necessity of taking into account the censoring mechanism. The extension of the underlying censoring model for the multinomial problem to the finite population sampling problem should lead methods for handling certain kinds of censored observations in finite population sampling.

As we have seen in Example 2 there is a relationship between the estimator $\delta_{m,w}$ discussed in Section 3.1.1 and the Dirichlet process prior (Ferguson, 1973). Let v_o be a sigma-additive non-null finite measure on the Borel sets of the real numbers. Let v_o be the parameter of a Dirichlet process for the unknown distribution function F, i.e. $F \in \mathcal{D}(v_o)$. Suppose F is chosen according to the Dirichlet process defined by v_o, and that given F, $Z_1, \ldots, Z_n$ are independent and identically distributed according to F. Then under this model the predictive distribution of a future observation Z_{n+1} given that $Z_1 = z_1, \ldots, Z_n = z_n$ is just

$$\frac{v_o(\mathcal{R})}{v_o(\mathcal{R}) + n} \frac{v_o}{v_o(\mathcal{R})} + \frac{n}{v_o(\mathcal{R}) + n} F_z \tag{3.31}$$

where F_z is the empirical distribution function of $z_1, \ldots, z_n$. Suppose now that $\mathbf{v} = (v_1, \ldots, v_k)^T$ is a probability distribution defined over $\mathbf{a} = (a_1, \ldots, a_k)^T$, a vector of distinct real numbers. Assume that this distribution is our prior guess about the shape of F. Let w be the weight we attach to our prior guess and let $\mathbf{v}_o = w\mathbf{v}$. Then the probability structure described in (3.31) is exactly the probability structure of the 'seeded' urn, described in Section 3.1.1 which led to the estimator $\delta_{m,w}$ where m is the mean of the probability distribution $\mathbf{v}$.

The estimator of Example 1 can be thought of as the limiting case as $v_o(\mathcal{R})$ tends to zero, i.e. the 'noninformative' Dirichlet prior. We shall now show that the sequence of priors used in

this example leads to the Bayesian bootstrap (Rubin, 1981). Consider again the nonparametric estimation problem described in the last section. Given $Z_1, \ldots, Z_n$ which are independent and identically distributed according to F, which is assumed to belong to some large 'nonparametric' set $\mathcal{F}$, we wish to estimate $\gamma(F)$. Let $z = (z_1, \ldots, z_n)^T$ be a typical vector of possible outcomes. Given z let $B(z) = \mathbf{b}$ be the vector of distinct values appearing in z, arranged in increasing order. Let $C(z) = \mathbf{c}$ be the count vector associated with $\mathbf{b}$, i.e. the ith coordinate of $\mathbf{c}$ is $c_z(i, n)$, the number of times b_i appears in the sample. If the members of z are all distinct then $\mathbf{b}$ is just the order statistic of z and $\mathbf{c}$ is a vector of one's.

We now give a noninformative Bayesian method for estimating $\gamma(F)$. Assume that the sample z has been observed and $B(z) = \mathbf{b}$ and $C(z) = \mathbf{c}$ are known. Next we assume that F must belong to the parameter space $\Lambda = \mathcal{F}(\mathbf{b})$ instead of the larger space $\mathcal{F}$. Furthermore we take as our 'prior distribution' over $\Lambda = \mathcal{F}(\mathbf{b})$ the Dirichlet distribution with parameter vector $\mathbf{c}$. This 'prior' in turn leads to a 'posterior distribution' for $\gamma(F)$ which can be used in the standard Bayesian ways to get point or interval estimators for $\gamma(F)$. Now this posterior is exactly the Bayesian bootstrap (Rubin, 1981). Moreover it is the 'posterior' that arises from the sequence of priors given in Example 1. Hence it has a stepwise Bayes justification and must lead to an admissible estimator for $\gamma(F)$. And finally, by (3.29), we see that an estimator which arises from the Polya posterior is just the finite population sampling analogue of the corresponding estimator based on the Bayesian bootstrap. Therefore the Polya posterior is just the Bayesian bootstrap adapted to the finite population sampling problem. In the next section we will see that the Bayesian bootstrap leads to improved estimators of population quantiles in nonparametric problems.

3.6.3 Estimating a quantile

In the previous section we saw that the Polya posterior was the finite population sampling analogue of the Bayesian bootstrap for nonparametric problems. In Section 2.8.2 we saw that for estimating the median of a finite population an estimator based on the Polya posterior seemed to be preferred to the sample median. These facts suggest that when estimating a quantile in a nonparametric problem an estimator based on the Bayesian bootstrap should be

preferred to the sample quantile. In this section we shall see that this is indeed the case. Now it might seem silly to choose, after the sample z has been observed, as our 'posterior' distribution one which is restricted to the subclass $\Lambda = \mathcal{F}(\mathbf{b})$ where $B(z) = \mathbf{b}$. As we have seen, under this 'posterior' the predictive distribution of an additional observation from the population is just the empirical distribution function. This gives zero probability to the event that a future observation is unequal to the observed values in the sample. It was just this fact that led Rubin to question the broad applicability of the Bayesian bootstrap. On the other hand, for a large sample size it does seem to give a sensible noninformative Bayesian approach to nonparametric problems.

Suppose $Z_1, Z_2, \ldots$ are real valued random observations where ties have probability zero. Hill (1968) introduced the assumption A_n which asserts that, conditional on the first n observations, the next observation Z_{n+1} is equally likely to fall in any of the $n+1$ open intervals defined by the successive order statistics in the sample. Furthermore if the sequence is exchangeable and λ_i is the proportion of future observations falling in the ith interval then A_n implies that conditional on the first n observations the distribution for the vector of λ_i is Dirichlet with a parameter vector of all ones. This argument is made under the assumption that the probability measures involved are countably additive. But then Hill shows that there is no countably additive probability distribution on the space of observations such that the conditional distributions will agree with A_n.

From some points of view A_n is a more attractive formulation for a noninformative nonparametric Bayesian approach than the Bayesian bootstrap. For one thing it does not assume that any future observation must equal one of the past observations. On the other hand even if one is willing to assume that given the first n observations the conditional distribution of the λ_i is Dirichlet with a parameter vector of all ones it is not clear how the probability assigned to each interval should be spread out over the interval. One obvious possibility is to distribute the probability uniformly over each interval. Banks (1988) called this approach the smooth Bayesian bootstrap and compared it to the usual Bayesian bootstrap and several other bootstrap methods as well. Based on simulation studies for interval estimators he declared the smooth Bayesian bootstrap the winner. In part this was based on the fact

that confidence intervals based on the smooth Bayesian bootstrap were slightly more accurate than those based on other methods.

As noted above, one theoretical justification for the Bayesian bootstrap is that it is a stepwise Bayes procedure and hence will lead to admissible procedures. Lane and Sudderth (1978) showed that Hill's A_n is consistent with a finitely additive probability measure on the space of observations. This in turn shows that A_n is coherent. (Coherence is a slightly weaker property than admissibility since a procedure may be coherent but not admissible.) This, however, is not a justification for the smooth Bayesian bootstrap since their work does not directly compute the necessary conditional distributions. In fact the results of Section 7 of their paper suggests that the smooth Bayesian bootstrap does not arise from some finitely additive prior distribution.

We now suggest another modification of the Bayesian bootstrap which is a smoother version of Rubin's Bayesian bootstrap and is also a stepwise Bayes procedure. We suppose that for every member of $\mathcal{F}$ its support is a subset of the finite interval $I = (\underline{d}, \overline{d})$. Let $g = (g_1, \ldots, g_k)$ be a grid defined on I, i.e. $\underline{d} = g_1 < \cdots < g_k = \overline{d}$. We assume that such a grid g is given and fixed. Let $\mathcal{F}(g)$ be the subset of $\mathcal{F}$ which contains all the density functions on I which are constant on each subinterval of I, defined by the grid g. After z is observed let $B_g(z) = \mathbf{b}_g$ be the subintervals of the grid g which contain at least one observation. Let $C_g(z) = \mathbf{c}_g$ be the count vector, i.e. the ith component of $\mathbf{c}_g$ is just the number of observations that fell in the ith occupied subinterval. Now given a sample z we will assume that our 'posterior' is concentrated on the subset of $\mathcal{F}(g)$ which consists of all those densities whose support is contained in $\mathbf{b}_g$. The 'posterior' distribution of the amount of probability assigned to each of these subintervals will just be Dirichlet with parameter vector $\mathbf{c}_g$. Then by assumption these 'posterior' probabilities are uniform over their respective subintervals. It is easy to check that this 'posterior' is in fact a stepwise Bayes procedure. The argument is exactly the same as the one used for the Bayesian bootstrap except it is now applied to subintervals defined by the grid rather than the actual values. The above procedure can be thought of as a specialization of Rubin's Bayesian bootstrap to categorical data, where the categories are just the subintervals of the grid. It assumes a histogram model with the probabilities assigned to the subintervals as the parameters. So in this sense it is not really nonparametric since the grid is determined before

the sample is chosen. The Banks procedure is similar in spirit except that he lets the subintervals of the grid be determined by the observed values in the sample. Although this is intuitively quite appealing, as we noted before there is no known Bayes or stepwise Bayes justification for a sample-dependent grid.

In what follows we will be comparing the point estimator, based on these three methods, along with the appropriate sample quantile, as estimators of a population quantile. We denote the Bayesian bootstrap by bbs, the smooth Bayesian bootstrap of Banks by sbbs, and the Bayesian bootstrap based on a grid by gbbs.

We begin by studying the usual sample quantile and an estimator based on Rubin's Bayesian bootstrap. Let $q = q(F)$ denote the qth quantile of F, the quantity to be estimated. Recall that given z we will take as a 'posterior distribution' for F a Dirichlet distribution with parameter vector $\mathbf{c}$ over the simplex $\Lambda = \mathcal{F}(\mathbf{b})$ when $B(z) = \mathbf{b}$ and $C(z) = \mathbf{c}$. This induces a 'posterior' for $q = q(F)$. If the loss function is squared error the corresponding estimator is the mean of this distribution. This estimator can only be found approximately by simulating the 'posterior distribution' of $q = q(F)$. For example, we could take R independent observations from the appropriate Dirichlet distribution. For each observation, compute the qth quantile and then find the mean of these R computed quantiles. We will call this estimator the Bayesian bootstrap estimator of $q = q(F)$ and is denoted by bbs. We begin by comparing it to the usual naive estimator of $q = q(F)$, i.e. the qth quantile of the sample, denoted by smp.

Since we are estimating population quantiles the most natural loss function is absolute error. For this loss function the Bayes estimator of $q = q(F)$ is just the median of our 'posterior distribution'. When estimating the population median Ferguson (1973) shows that the median of our 'posterior distribution' is just the median of z, i.e. the sample median. For other quantiles this estimator cannot be found explicitly. In some cases it behaves quite similarly to the estimator smp while in other instances it is outperformed by both the bbs and smp estimators. For this reason we will only compare the estimators bbs and smp, but using absolute error as our loss function.

The two estimators were compared for six different populations. The first two populations were a beta(20, 20) and a beta(0.7, 0.8) denoted by Beta1 and Beta2. The remaining four were Cauchy with location and scale parameters 0 and 1, the standard exponential,

the gamma with shape parameter 20 and scale parameter 1, and the lognormal with mean and standard deviation (of the log) 4.9 and 0.586. For each population we considered samples of size 11, 25 and 50. In each case we estimated the tenth, twenty-fifth, fiftieth, seventy-fifth and ninetieth quantiles. These quantiles are denoted by $q10$, $q25$, $q50$, $q75$ and $q90$. In each case we observed 500 random samples from each population. Given a sample we computed bbs by taking $R = 500$ observations from the appropriate Dirichlet distribution. These simulations were done using S. When computing the quantile of a sample of size n, S linearly interpolates between order statistics of the sample, assuming the ith order statistic is the $(i-0.5)/n$ quantile. The results of the simulations are found in Table 3.9. The table gives the ratio of the average absolute error of the estimator mpq to the average absolute error of the estimator bbs.

We see from Table 3.9 that, except for the Cauchy distribution, the Bayesian bootstrap estimator performs better than the usual naive estimator in the vast majority of cases. The usual estimator seems to do better for extreme quantiles in the heavy tail of the distribution, especially for small sample sizes. The average amount of improvement for bbs over smp is somewhere between 5% and 10% percent although it can reach as high as 20%. Similar results hold for estimating $q90 - q10$ and $q75 - q25$ (Meeden, 1993). This suggests that for most nonparametric problems for which moments are assumed to exist one should use the Bayesian bootstrap not a sample quantile.

We will now present one possible explanation for this, perhaps somewhat surprising, fact. A holistic approach to nonparametric estimation suggests that one should first find a reasonable estimate of the entire population and then use this to find estimates of particular aspects of the population of interest. In some sense both the sample mean and sample median, as estimators of the population mean and median, seem to follow from this holistic approach. However, the sample mean makes fuller use of the information in the sample than does the sample median. It is just this fact that makes the sample median more robust than the sample mean. It has long been recognized that when estimating a quantile the usual naive estimator can be improved upon if some additional information about the population is known. For example when estimating the median of a symmetric population a symmetric pair of order statistics can be preferred to the sample median. The difficulty for

Table 3.9 *The ratio of the average absolute error for the estimator smp to the average absolute error for the estimator bbs.*

Population	Sample size	$q10$	$q25$	$q50$	$q75$	$q90$
	11	1.00	1.07	1.10	1.09	1.03
Beta1	25	1.18	1.10	1.08	1.12	1.13
	50	1.09	1.10	1.06	1.05	1.10
	11	0.75	1.00	1.23	1.10	0.81
Beta2	25	0.94	1.03	1.16	1.12	0.93
	50	0.94	1.07	1.09	1.07	1.01
	11	1.16	0.44	0.84	0.39	1.19
Cauchy	25	0.40	0.66	1.01	0.82	0.41
	50	0.62	0.98	1.02	0.97	0.63
	11	0.80	0.95	1.06	1.09	1.04
Exponential	25	0.97	1.02	1.07	1.08	1.12
	50	1.01	1.04	1.07	1.06	1.08
	11	1.02	1.07	1.12	1.09	1.05
Gamma	25	1.08	1.08	1.11	1.07	1.14
	50	1.09	1.11	1.07	1.09	1.08
	11	0.90	1.02	1.12	1.07	1.13
Lognormal	25	1.09	1.06	1.06	1.09	1.13
	50	1.07	1.04	1.04	1.04	1.07

the general nonparametric problem, when only vague prior information about the population is available, is how to improve upon the usual naive estimator. What is surprising is that the 'posterior' leading to the Bayesian bootstrap seems to do just that in essentially an automatic manner.

Often one is interested in interval estimators as well as point estimators. As we have seen in finite population sampling, the Polya posterior, which is the Bayesian bootstrap adapted to finite population sampling, gave an interval estimator of the finite population

median which was very close to a frequentist interval. For the nonparametric problem of estimating a median, the Bayesian bootstrap intervals are asymptotically equivalent to the usual nonparametric intervals based on the order statistic (Efron, 1982). Note that since this interval is based on two extreme order statistics it makes more use of the information in the sample than a single sample quantile does.

We now wish to compare the three Bayesian bootstrap methods bbs, sbbs and gbbs. In each case when computing a point estimate for a quantile we will take the mean of the simulated values under the 'pseudo posterior'. That is, we will do the same for each method here as we did when computing the estimator bbs earlier. We did this using the gamma distribution with shape parameter 20 and scale parameter 1. We also consider three different grids for gbbs. One grid just went from 0 to 60 with each step having length 1. For this distribution 8.96 is the 0.001 quantile while 36.7 is the .999 quantile. A second grid started with the values 0, 2, 4, 5, 6, 7, 8 and 8.5. It then divided (8.96, 36.7) into 100 successive subintervals of equal probability. It then ended with the values 40, 45, 50, 55 and 60. The shortest subinterval in this grid was just slightly larger than 0.11. The third grid went from 0 to 8 in steps of 1, then from 8.5 to 30 in steps of 0.1 and from 30 to 60 in steps of 1. The simulation results for the three grids were very similar and hence we will present only the results for the third grid. So the choice of a grid does not seem to matter very much as long as there are no large subintervals which contain lots of probability.

To use the smooth Bayesian bootstrap one needs to assume that the support of the distribution is confined to some known interval. We took the lower bound to be 0. Although technically correct this is not a good idea because the subinterval defined by 0 and the minimum of the sample is too large and contains a region which is essentially assigned probability zero. This may not matter much when estimating the median but could have some effect when estimating smaller quantiles. For the upper end we assumed that the last interval was just (max, max + 1) where max was the maximum value of the sample. Some comparisons are given in Tables 3.10 and 3.11.

Overall the performance of the three 'Bayesian bootstrap' estimators is quite similar. For a small sample size the smooth Bayesian bootstrap seems to give more accurate coverage probabilities but as expected can be sensitive to the assumed bounds for the distribu-

Table 3.10 *The average value and absolute error of the point estimator for 500 random samples from a gamma distribution for three 'Bayesian bootstrap' methods.*

Quantile	Sample size	Estimator	Average value	Average absolute error
		bbs	17.21	1.14
	11	sbbs	16.81	1.08
		gbbs	17.21	1.14
$q25 = 16.83$				
		bbs	16.96	0.77
	25	sbbs	16.82	0.77
		gbbs	16.96	0.77
		bbs	19.89	1.18
	11	sbbs	19.89	1.18
		gbbs	19.89	1.18
$q50 = 19.67$				
		bbs	19.74	0.79
	25	sbbs	19.75	0.80
		gbbs	19.75	0.79
		bbs	22.68	1.37
	11	sbbs	23.08	1.43
		gbbs	22.68	1.36
$q75 = 22.81$				
		bbs	22.83	0.95
	25	sbbs	23.00	0.96
		gbbs	22.83	0.95

tion. Now it is intuitively clear that Rubin's Bayesian bootstrap, for small sample sizes, will under-estimate the amount of variation in the population. Hence Bank's smooth Bayesian bootstrap which spreads out the 'posterior probability' could be an improvement. However, to call it a smooth Bayesian bootstrap seems to be a misnomer since there does not appear to be any Bayesian justification for the method. The modification of Rubin's Bayesian bootstrap when the sample space is partitioned into a grid has the intuitive appealing property that the predictive distribution for a

Table 3.11 *The average length and relative frequency of coverage for a 0.95 credible interval estimator for 500 random samples from a gamma distribution for three 'Bayesian bootstrap' methods.*

Quantile	Sample size	Estimate	Average length	Relative frequency of coverage
		bbs	6.31	0.916
	11	sbbs	7.83	0.970
		gbbs	6.26	0.916
$q25$				
		bbs	4.30	0.944
	25	sbbs	4.39	0.956
		gbbs	4.31	0.948
		bbs	6.36	0.916
	11	sbbs	6.68	0.936
		gbbs	6.33	0.912
$q50$				
		bbs	4.41	0.950
	25	sbbs	4.39	0.960
		gbbs	4.37	0.954
		bbs	7.92	0.904
	11	sbbs	7.98	0.914
		gbbs	7.81	0.900
$q75$				
		bbs	5.46	0.966
	25	sbbs	5.57	0.964
		gbbs	5.42	0.962

future observation is no longer concentrated on the observed values. However, point and interval estimates based on it seem to differ little from those based on Rubin's Bayesian bootstrap. Moreover for moderate and larger sample sizes, based on the simulations given here, there seems to be little difference between the three methods.

Up until now most estimation studies have been concerned with interval estimation and obtaining estimates of variance. Although such problems are important in practice, the problem of point estimation is of interest as well. The main purpose of this section was

to demonstrate that estimates of population quantiles based on Rubin's Bayesian bootstrap are preferred to the usual naive estimates. Moreover this method yields a sensible noninformative Bayesian approach to these problems. For a more informative Bayesian approach to such problems see Doss (1985). Yang (1985) presents another approach to the nonparametric estimation of quantiles. It is essentially a kernel-type estimator. The optimal choice of the window width depends both on the sample size and the underlying distribution. He suggests a bootstrap technique to estimate the optimal width. Since this involves doing many different bootstraps for each sample it was not included in this study.

3.7 Linear interpolation

In Chapter 2 and the earlier sections of this chapter we considered various situations where the Polya posterior or modifications of it led to sensible inferences. In all such problems, even though no prior distribution needed to be specified, it was necessary for the statistician to believe that the characteristic of interest for the units, the y_i's, or some function of the y_i's, were roughly exchangeable. In this section we consider some cases where ones's prior beliefs about the units cannot be characterized by some type of exchangeability. For example, suppose it is reasonable to assume that units whose labels are close together are more alike than units whose labels are far apart. This may arise naturally from the geometry of the problem. In other cases the statistician may use all of his or her prior information to relabel the population so that this condition is satisfied. How should one make use of this state of affairs when choosing an estimator? One possibility is to stratify the population. The strata would consist of units whose labels are a set of successive integers. Even if it were not practical to make the strata small enough to be nearly homogeneous, it would still be foolish to ignore this information. For example, suppose in our sample we have observed that unit i_1 has the value a and unit i_2 has the value b where $i_1 < i_2$ and none of the units whose labels fall between i_1 and i_2 belong to the sample. Let j denote the label of a typical unsampled unit with $i_1 < j < i_2$. If we really believed that units whose labels are close together are more alike than units whose labels are far apart, then our estimate for the value of the unit with label j should depend only on the numbers a and b, that is, on the values of the two units in the sample which are closest to it. Moreover, a natural choice

for this estimate is $[(j-i_1)b+(i_2-j)a]/(i_2-i_1)$, that is, just the value at j of the straight line passing through the points (i_1,a) and (i_2,b). That is, for every unsampled member of the population, we take as our estimate the value of the linear interpolation between its two closest points in the sample. This in a straightforward way leads to an estimate of the population total or population mean. In Section 3.7.1 we show that this estimator has a stepwise Bayes justification and is admissible under squared error loss. In Section 3.7.2 some examples will be considered and, in addition, the related interval estimators studied.

3.7.1 Admissibility

In this section we will use a stepwise Bayes argument to prove the admissibility of the linear interpolation estimator. This estimator is appropriate when the statistician believes that units with labels close together are more alike than units with labels that are far apart. For an unsampled unit the estimate of its value is just the linear interpolation of the unobserved values of the two units which are closest to it in the sample. For units with labels less than the smallest label appearing in the sample, their estimate is just the observed value of the unit with the smallest label in the sample. Units with labels larger than the largest label appearing in the sample are handled in a similar manner.

Let δ^{li} denote this linear interpolation estimator of the population mean. For a typical data point $z=(s,z_s)$ where $n(s)=n$ and $s=\{i_1,\ldots,i_n\}$ and with $1\le i_1<\cdots<i_n\le N$, where N is the population size, it is easy to give an explicit expression for $\delta^{li}(z)$. The cases $n=1$ and $n=N$ are trivial. In the case $n=2$

$$\delta^{li}(z)=\frac{1}{2N}\{z_{i_1}(i_1+i_2-1)+z_{i_2}(2N-i_1-i_2+1)\} \tag{3.32}$$

and for the case $2<n<N$

$$\begin{aligned}\delta^{li}(z) &= \frac{1}{2N}\Bigg\{z_{i_1}(i_1+i_2-1)\\ &+ \sum_{j=2}^{n-1} z_{i_j}(i_{j+1}-i_{j-1})+z_{i_n}(2N-i_{n-1}-i_n+1)\Bigg\}.\end{aligned} \tag{3.33}$$

Note that in the special case when all the members of the observed

data point have the same value, the value of δ^{li} is just this common value.

Theorem 3.6 *Let* $\mathbf{b} = (b_1, \ldots, b_k)^T$ *be a vector of distinct real numbers. Then for estimating the population mean, under squared error loss, the estimator* δ^{li} *is admissible, under any design* p*, when the parameter space is* $\mathcal{Y}(\mathbf{b})$*, defined in equation* (1.1).

Proof. The admissibility of δ^{li} will be demonstrated by showing that it is a stepwise Bayes estimator against a sequence of priors defined on $\mathcal{Y}(\mathbf{b})$. If $k = 1$ the parameter space contains just one point and the result is obvious. So we suppose that $k \geq 2$.

Our first prior puts mass $1/k$ on the k points of the form $\mathbf{y} = (b_j, \ldots, b_j)^T$ for $j = 1, \ldots, k$. Under this prior the only points in the sample space which receive positive probability are those where the units in the observed data point are all equal to one of the b_j's. It is easy to check that under this prior δ^{li} is the unique Bayes estimate for all such outcomes.

Thus we have shown that δ^{li} is Bayes against our first prior. We will next account for all those data points where the units in the sample take on only two possible values. For the rest of the proof, p will be a fixed design which assigns positive probability to at least one sample s with $n(s) \geq 2$.

For notational convenience, let $a = b_1$ and $b = b_2$. In the next stage we will define a sequence of priors on disjoint subsets of $\mathcal{Y}((a, b)^T)$. This will allow us to show that δ^{li} is a stepwise Bayes estimator for all possible data points where the observed values are either a or b.

We now let

$$\mathcal{Y}^1((a,b)^T) =$$
$$\{\mathbf{y} : y_i = a \text{ for } i = 1,\ldots,j \text{ and } y_i = b \text{ for } i = j+1,\ldots,N$$
$$\text{or } y_i = b \text{ for } i = 1,\ldots,j \text{ and } y_i = a \text{ for } i = j+1,\ldots,N$$
$$\text{for some } j = 1,\ldots,N-1\}.$$

On this subset of $\mathcal{Y}((a,b)^T)$ we take as our prior π the distribution which assigns equal probability to each point of $\mathcal{Y}^1((a,b)^T)$. Under this prior the data points which are possible now and were not taken care of in the first stage are all those outcomes where the units are first all a's and then followed by all b's or vice versa. Let $z = (s, z_s)$ be such a data point with $n(s) = n$. For this step we must consider two cases.

We first assume that there exists an integer j^* such that $1 \leq j^* \leq n-1$, $i_{j^*} + 1 = i_{j^*+1}$ and

$$z_{i_j} = \begin{cases} a & \text{for } j = 1,\ldots,j^* \\ b & \text{for } j = j^*+1,\ldots,n. \end{cases}$$

For such a z the posterior probability distribution under π concentrates all its mass at one point of $\mathcal{Y}((a,b)^T)$ and the resulting Bayes estimate is just $\delta^{li}(z)$. Note that a similar result holds if a and b are interchanged.

Next we consider the case of a z such that j^* is as above except now $i_{j^*} + 1 < i_{j^*+1}$. For notational convenience, set $i_o = i_{j^*}$ and $d = i_{j^*+1} - i_{j^*}$. For such a z and the prior π the posterior assigns equal probability to the d points of the form

$$\mathbf{y} = (\underbrace{a,\ldots,a}_{i_o+r}, \underbrace{b,\ldots,b}_{N-i_o-r})^T$$

for $r = 0,1,\ldots,d-1$. Hence for $r = 1,\ldots,d-1$, the posterior probability that $y_{i_o+r} = b$ is r/d and so

$$\begin{aligned} \sum_{r=1}^{d-1} E(y_{i_o+r} \mid z) &= \sum_{r=1}^{d-1} \left(\frac{d-r}{d}a + \frac{r}{d}b\right) \\ &= (d-1)(a+b)/2. \end{aligned} \tag{3.34}$$

Now the posterior expectation of any other unobserved unit must be either a if its label is less than i_{j^*} or b if its label is greater than i_{j^*+1}. It is easy to check that this fact along with (3.34) shows for

such a z the resulting Bayes estimate is just $\delta^{li}(z)$. Again a similar result holds if a and b are interchanged.

If the design p only assigns positive probability to samples of size two or less, the proof of the theorem would be completed at this point. Hence we will assume that this is not so. In this case it might be possible to observe a sample of size three of larger where $y_{i_1} = y_{i_3} = a$ and $y_{i_2} = b$. The next stage of the argument will consider points of this type.

Let

$$\mathcal{Y}^2((a,b)^T) = \{\mathbf{y} : \text{there exist integers } 1 \le j_1 < j_2 < N \text{ such that } y_i = a \text{ for } 1 \le i \le j_1 \text{ and } j_2 < i \le N \text{ and } y_i = b \text{ for } j_i < i \le j_2 \text{ or vice versa}\}.$$

Just as before we take as our prior π, the distribution which assigns equal mass to the points in $\mathcal{Y}^2((a,b)^T)$. Next we need to compute the Bayes estimate against π for all data points which receive positive probability under this prior and which were not accounted for by the previous priors. It is easy to check that for such data points the resulting Bayes estimator agrees with δ^{li}.

The argument now proceeds in a similar fashion where at any step, the prior is the uniform distribution over the appropriately chosen subset of the parameter space, and given a data point, z, the posterior distribution will give equal weight to all those $\mathbf{y}$'s which are the simplest possible step functions consistent with z. To determine the proper order, we introduce a definition for $\mathbf{y} \in \mathcal{Y}(\mathbf{b})$. We say that $\mathbf{y} = (y_1, \ldots, y_N)^T$ has **order** r if the number of labels i such that $y_{i+1} \neq y_i$ is r. Note that for a sample s the order of $y(s)$ is well defined, where we take the order of $y(s)$ to be 0 if s contains just one point. Hence $\mathcal{Y}^2((a,b)^T)$ is just all those vectors of order 2 whose individual components are either a or b.

The next step is to consider

$$\mathcal{Y}^3((a,b)^T) = \{\mathbf{y} : \mathbf{y} \in \mathcal{Y}((a,b)^T) \text{ and } \mathbf{y} \text{ is of order } 3\}.$$

The next step would be to consider $\mathcal{Y}^4((a,b)^T)$ and continue in this way until we have accounted for all possible data points in which just a and b appear. The value of the highest order will depend on p. It will be

$$n_o = \max\{n(s) : p(s) > 0\} - 1.$$

The next step is to repeat the process for $\mathcal{Y}((b_1,b_3)^T)$ and then

for $\mathcal{Y}((b_1, b_4)^T)$ and so on to $\mathcal{Y}((b_{k-1}, b_k)^T)$. In this way we will take care of all data points where just two different values appear. (Recall in the first stage we took care of all sample points where just one value appears.) In the next stage we will take care of all data points where just three values appear. In the following stage we will take care of all data points where just four values appear and so on. If $k \leq n_o$, the final stage begins by defining $\mathcal{Y}^{k-1}(\mathbf{b})$ and proceeding as before. If $k > n_o$, the process stops after all data points in which n_o different values appear have been considered. In either case this completes the proof since we have demonstrated that δ^{li} is a unique stepwise Bayes estimator. □

Because the theorem holds for every possible choice for the vector $\mathbf{b}$, we have as a corollary that δ^{li} is admissible when the parameter space is $\mathcal{R}^N$.

Given that one will use the estimator δ^{li}, how should one choose an appropriate design? For example, suppose we can use any design in the class of designs of fixed sample size n, where n is some fixed positive integer. We can use the approach described in Section 3.4 to find a design such that the resulting design and the estimator δ^{li} is uniformly admissible for this problem. This was discussed in Meeden (1992). There it was seen that one should choose a nonrandom design which concentrates its mass on a sample s which is symmetric about $N/2$ and whose successive differences are nearly equal; the only surprise being that the smallest and largest members seemed to be quite far from their respective endpoints of the population. For example, when $N = 100$ and $n = 11$ it was found that one should choose any s with $i_1 = 18$ and $i_{11} = 83$ and five differences equal to six and five differences equal to seven.

The above suggests that when using the estimator δ^{li}, one only needs to consider nonrandom samples whose labels are approximately equally spaced throughout the population. In some cases, however, it may be desirable to choose a sample whose labels are not equally spaced. A statistician could choose fewer units in regions where it was believed the units were quite similar and more units in regions with more variability. For example, it may be the case that y_i is approximately an increasing function of i; however, in some regions it is believed to be nearly constant while in other regions it increases quite rapidly. Formally, one could divide the population into several regions and use the estimator δ^{li} within each region separately. More informally, one could choose the sam-

pled units to reflect prior beliefs about variability within various regions of the population. In particular, one may not wish to choose the first and last units of the sample so far from the population endpoints as the theory suggests. This would be the case when one's prior knowledge about units close to the ends is not as reliable as units closer to the centre of the population.

Presumably either the estimator δ^{li} or others very similar to it have been considered in the past. However, we do not know of any explicit references to this estimator. If the labels appearing in the sample are nearly equally spaced throughout the population, then this estimator will be approximately equal to the sample mean. However, when the labels appearing in the sample are not equally spaced, the estimator can be quite different from the unusual estimator.

3.7.2 Some examples

In this section we will consider some examples to see how the linear interpolation estimator works in practice. In the last section we saw that this point estimator is in fact a stepwise Bayes estimator which arises from a sequence of priors which is suitable when the statistician believes that units whose labels are close together are more alike than units whose labels are far apart. Just as with the Polya posterior the resulting posterior can be used to find interval estimates as well as point estimates of parameters of interest. We will simulate observations from this posterior to find approximately the 'posterior' distribution of the population mean. As before (see Section 2.7) we let $q025$ and $q975$ be the 0.025 quantile and 0.975 quantile, respectively, of this set of simulated values for the population mean. Then $(q025, q975)$ will be our announced credible set with 'posterior probability' 0.95. One would expect that for parameter points $\mathbf{y}$ which are sufficiently smooth, that is, for populations whose members, with labels close together, are more alike than members with labels far apart, its coverage probability should be about 0.95. While for less smooth parameter points, the coverage probability could be considerably less. In what follows, this stepwise Bayes procedure will be compared to frequentist procedures for various populations.

Jessen (1978) on pages 18 and 19 discussed the following experiment due to F. Yates. The population was a collection of 126 stones of various sizes displayed on a table. Selectors were asked

to purposely pick representative samples of various sizes. Then the estimates of the population mean based on their samples were compared to estimates based on random samples of the same size. It was found that selectors did better than random sampling if the sample size was less than eight while random sampling did better for larger sample sizes. In some sense this is quite a surprising result and suggests how difficult it can be to attempt to incorporate prior information in a naive fashion.

One approach to this problem is to use the techniques of the previous section. For example, one could to attempt to arrange the stones in increasing order from the smallest to the largest and then use the linear interpolation estimator. If the ordering is reasonably accurate, then the linear interpolation estimator should perform better than the usual estimator based on random sampling.

A frequentist sampler, faced with a population which was known to be smooth, would not use simple random sampling but would stratify. In the examples that follow, all the populations will be stratified. A sample will be drawn by choosing one unit at random from each stratum. A 95% confidence interval will be constructed using the method of collapsed strata; see Cochran (1977), page 139. This is done by assuming the usual estimate is approximately normally distributed and getting an estimate of its variance by pairing successive strata. (Cochran notes that in some cases this can lead to an overestimate of the sampling variance of the estimator.) This will be called the ST interval and will be compared to the approximate 0.95 credible interval based on the 'posterior' which leads to the linear interpolation estimator. This will be called the LI interval and the two methods will be compared for six different populations.

The first population, pp1, was constructed from the set $A = \{a(1), \ldots, a(100)\}$ of real numbers. $a(\cdot)$ was a piecewise linear function consisting of seven different pieces. The change points were after the points 6, 16, 37, 66, 86 and 96. Then y_i was set equal to $a(i)$ plus a realization of a normal random variable with mean 0 and variance $(i+3)/4$. In the first plot of Figure 3.11, pp1 is plotted against its labels. Note that pp1 is reasonably smooth. We then divided pp1 into 11 strata. The first stratum contained the first five units, while the last contained the last five units and the remaining nine strata each contained ten successive units. Five hundred stratified samples were taken where one unit was taken at random from each stratum. The results are given in the first row of Table 3.12.

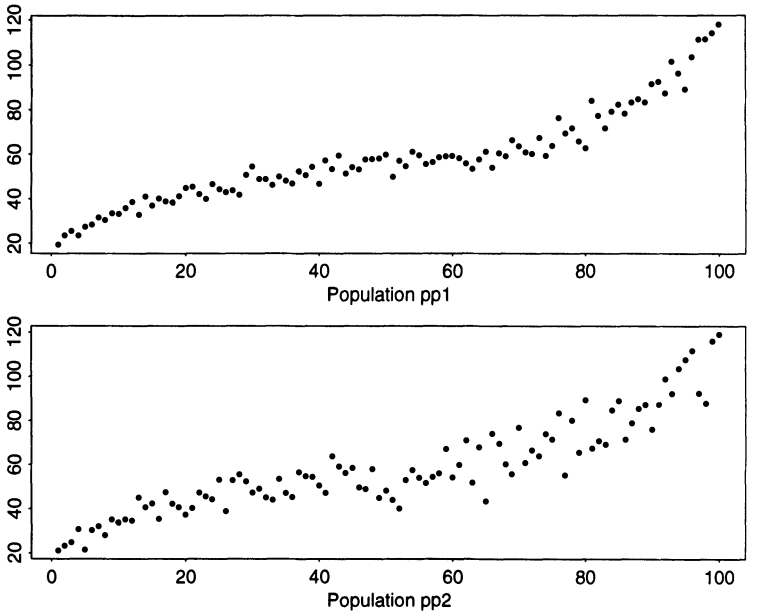

Figure 3.11 *Plots of the populations pp1 and pp2.*

We see, that on average, the ST interval is almost twice as long as the LI interval. Furthermore, it covers almost all the time while the LI interval's frequency of coverage is exactly 0.95 in this case. In retrospect, this is not too surprising since the ST interval makes no use of the smooth nature of pp1 beyond the stratification, while the LI interval is based on the smoothness assumption.

A second population, pp2, was constructed from the set A. This time y_i was set equal to $a(i)$ plus a realization of a normal random variable with mean 0 and variance $(i+3)$. This population is plotted in the second plot of Figure 3.11 and is not nearly as smooth as pp1. It was stratified in the same way as pp1 and 200 stratified samples were taken. The results are in the second row of Table 3.12. As was to be expected the LI interval does poorly since pp2 is not sufficiently smooth.

The final four populations come from sunspot data discussed in Andrews and Herzberg (1985). These data are successive monthly means of daily sunspot numbers beginning in the year 1749. Because of the cyclic nature of the sunspot activity, we would expect

this sequence of observations to exhibit some smoothness. The population ppssp1 is the first 100 observations of this sequence, ppssp5 the first 500, ppssp10 the first 1,000 and ppssp24 the first 2400. We stratified ppssp1 in the same manner as pp1 and pp2, while the last three populations were divided into 25, 50 and 120 strata, respectively, where each stratum was of length 20. For each of these four populations, 500 random samples of one observation from each stratum were taken and the two intervals computed. The results are given in the last four lines of Table 3.12. We see that the ST interval is too long and overcovers. On the other hand the LI interval is considerably shorter but yields the appropriate frequency of coverage only when the population gets sufficiently large. This indicates that the degree of smoothness of a population depends in part on how many units are in the population.

In the chapter on systematic sampling, Cochran (1977) discussed approaches to estimation for some naturally occurring populations like the sunspot data. In such cases it is often very difficult to get a sensible estimate of variance and usually such estimates depend on an assumed model for the data. Note that to use the LI interval one only needs to assume that the population is smooth and have some idea about the length of the sunspot cycles. This is important since the sunspot cycles need to be much longer than the sampling strata, so that the population is approximately linear between observations.

We chose the ST interval as the one to compare to the LI interval because we thought it would be the strongest competitor. For another possibility, consider population ppssp24 and divide it into 60 strata each of size 40 and consider the design which selects two units at random without replacement from each stratum. Five hundred samples were taken and the usual interval computed. The average length was 7.98 and 0.928 was the relative frequency of coverage. This is not as good as the LI interval discussed above. This is not surprising since, on the average, choosing two observations out of every 40 units should not be as representative as choosing one out of every 20.

The LI interval arises from an intuitively appealing pseudo posterior distribution. If one believes that the population is smooth, then, given the sample, this posterior puts equal mass over the set of parameter points which are as smooth as possible and still are consistent with the observed data point. Since this is true no matter what the actual values appearing in the data point are, the

Table 3.12 *Comparison of the 95% ST and the 0.95 LI intervals for the population mean for six populations with 500 samples, except pp2 which had 200 samples.*

	Average length of ST interval	Relative frequency of coverage	Average length of LI interval	Relative frequency of coverage
pp1	9.12	0.99	4.61	0.95
pp2	11.35	0.92	5.21	0.71
ppssp1	21.64	0.99	8.74	0.72
ppssp5	28.61	0.99	12.59	0.90
ppssp10	15.95	1.00	7.08	0.92
ppssp24	9.65	0.99	5.02	0.95

estimator should perform well over a wide class of smooth populations. We believe that the results of Table 3.12 indicate that this is so. Dealing with smooth populations is difficult not only from the straightforward Bayesian point of view but from the frequentist point of view as well. Beyond stratifying and taking a random sample of size 1 or 2 from each stratum, it is not clear how a frequentist would proceed and it is just these situations where the frequentist theory is the most unsatisfying. Selecting the labels of a finite population to make it smooth can be thought of as a generalization of stratification, that is, a non-Bayesian method of incorporating prior information. However, once the sample has been observed, one can use the pseudo posterior distribution to find point and set estimates. If the population is indeed smooth, then the set estimator will have good frequentist properties as well.

CHAPTER 4

Empirical Bayes estimation

4.1 Introduction

Empirical and hierarchical Bayes methods are becoming increasingly popular in statistics, especially in the context of simultaneous estimation of several parameters. For example, agencies of the Federal Government have been involved in obtaining estimates of per capita income, unemployment rates, crop yields and so forth for many state and local government areas. In such situations, quite often estimates of certain area means, or simultaneous estimates of several area means, can be improved by incorporating information from similar neighbouring areas. Estimation problems of this type are usually referred to as small area estimation problems.

The term 'small area' is commonly used to denote a small geographical area, such as a county, a municipality or a census division. It may also describe a 'small domain', that is a small subpopulation of people within a large geographical area defined by characteristics such as age, sex and race.

In recent years, there has been a growing demand for reliable small area statistics from both public and private sectors. These days, there is increasing government concern with issues of distribution, equity and disparity. For example, there may exist geographical subgroups within a given population that are far below the average in certain respects and need definite upgrading. Before taking remedial action, there is a need to identify such regions and, accordingly, one must have statistical data at the relevant geographical levels. Small area statistics are also needed for the apportionment of government funds, and in regional and city planning. In addition, there are demands from the private sector, since the policy making of many businesses and industries depends on local socio-economic conditions.

In the present and in the next chapter, we will mostly discuss Bayesian methods which among other things will be appropriate for small area estimation. Empirical and hierarchical Bayes meth-

ods are particularly well suited to meet this need. As described by Berger (1985), an empirical Bayes scenario is one in which known relationships among the coordinates of a parameter vector, say $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$, allow use of the data to estimate some features of the distribution. For example, one may have reason to believe that the θ_i's are iid from a prior $\pi_0(\lambda)$, where π_0 is structurally known except possibly for some unknown parameter λ. A parametric (EB) procedure is one where λ is estimated from the marginal distribution of the observations.

Closely related to the EB procedure is the hierarchical Bayes (HB) procedure which models the prior distribution in stages. In the first stage, conditional on $\Lambda = \lambda$, θ_i's are iid with a prior $\pi_0(\lambda)$. In the second stage, a prior distribution (often improper) is assigned to Λ. This is an example of a two-stage prior. The idea can be generalized to multistage priors, but that will not usually be pursued in this book.

It is apparent that both the EB and the HB procedures recognize the uncertainty in the prior information, but whereas the HB procedure models the uncertainty in the prior information by assigning a distribution (often noninformative or improper) to the prior parameters (usually called hyperparameters), the EB procedure attempts to estimate the unknown hyperparameters, typically by some classical method like the method of moments, the method of maximum likelihood etc., and use the resulting estimated priors for inferential purposes. It turns out that the two methods can quite often lead to comparable results, especially in the context of point estimation. This will be revealed in some of the examples appearing in the later sections. However, when it comes to the question of measuring the standard errors associated with these estimators, the HB method has a clear edge over a naive EB method. Whereas, there are no clearcut measures of standard errors associated with the EB point estimators, the same is not true with HB estimators. To be precise, if one estimates the parameter of interest by its posterior mean, then a very natural estimate of the risk associated with this estimator is its posterior variance. Estimates of the standard errors associated with EB point estimators usually need an 'ingenious approximation' as given, for example, in Morris (1983a, 1983b), whereas the posterior variances, though often complicated, can be found exactly.

The present chapter will deal with EB procedures, while the following chapter will deal with HB methods. The organization

of the remainder of the sections in this chapter is as follows. In Section 4.2, it is shown how some of the estimators derived in the earlier chapters as stepwise Bayes estimators, can also be derived using an EB approach. In Section 4.3, a simple EB model without covariates is introduced. This model is described in Ghosh and Meeden (1986). The robustness of the model without the normality assumption is discussed in Section 4.4 based on Ghosh and Lahiri (1987). Multistage extensions of the results of Section 4.4 are given in Section 4.5. Another generalization of the model given in Section 4.4 which includes covariates is discussed in Section 4.6. The results of this section unify in the context of a finite population, the work of Fay and Herriot (1979), Prasad and Rao (1990), and others. A slightly different model employed by Battese, Harter, and Fuller (1988) for EB analysis is discussed in Section 4.7.

4.2 Stepwise Bayes estimators

We have found in Chapters 2 and 3 a variety of admissible stepwise Bayes estimators of the finite population mean. In this section, we shall arrive at some of these estimators via an EB approach. An alternative derivation of these estimators via the HB method will be given in the next chapter.

Consider a Bayesian or superpopulation model for the population under which conditional on θ, we have that $y_1, \ldots, y_N$ are independently distributed with $E(y_i) = \theta a_i$ and $V(y_i) = \sigma_i^2$. The a_i and σ_i^2 are known constants. Then the finite population mean μ is estimated as

$$E(\mu|z) = \sum_{i \in s} z_i + \theta \sum_{j \notin s} a_j. \tag{4.1}$$

The EB estimator of μ is now obtained by estimating the unknown parameter θ. The weighted least squares estimator of θ is $\hat{\theta}_{LE} = \sum_{i \in s} a_i \sigma_i^{-2} z_i / \sum_{i \in s} a_i^2 \sigma_i^{-2}$. This results in the EB estimator of μ as

$$\hat{\mu}_{EB} = N^{-1} \left[n\bar{z}_s + \left(\sum_{i \in s} a_i \sigma_i^{-2} z_i / \sum_{i \in s} a_i^2 \sigma_i^{-2} \right) \sum_{j \notin s} a_j \right]. \tag{4.2}$$

The present EB approach is closely akin to Royall's prediction approach in finite population sampling. Like Royall's, our objective is to predict the unobserved units in the population based on

the observed units in the sample. We may note here that without additional assumption about the distribution of the y_i, it is not apparent how to construct credible (or predictive) sets for functions of $y_1 \ldots, y_N$ such as μ and other parameters. One may use the method of this section or use the HB approach discussed in (5.2).

Note that a typical estimator of the population mean as given in Royall (1970) (also Royall, 1971) is of the form

$$\hat{\mu}_R = N^{-1}\left[n\bar{z}_s + \left(\sum_{i\in s} v^{-1}(x_i)x_i z_i / \sum_{i\in s} v^{-1}(x_i)x_i\right)\sum_{j\notin s} x_j\right] \tag{4.3}$$

The estimator given in (4.3) is clearly obtainable from (4.2) with $a_i = x_i$ and $\sigma_i^2 = \sigma^2 v(x_i)$, $i = 1, \ldots, N$. The special case $v(x_i) = x_i$ leads to the ratio estimator $(\sum_{i\in s} z_i / \sum_{i\in s} x_i)\sum_{i=1}^N x_i$, while the choice $v(x_i) = x_i^2$ provides Basu's (1971) estimator. The admissibility of the latter is proved in Chapter 3.

Other interesting estimators are also obtainable from (4.2). For instance, putting $a_i = \pi_i$ and $\sigma_i^2 = \pi_i^2/(1-\pi_i)$, when we have $\sum_{i=1}^N \pi_i = n$, one obtains the celebrated Horvitz–Thompson estimator $\sum_{i\in s}(z_i/\pi_i)$ of μ. Many other estimators proposed in Vardeman and Meeden (1983), , are also obtainable as EB estimators.

4.3 Estimation of stratum means

This section considers simultaneous estimation of means from several strata when only a few observations are available from each individual stratum. Our results find application in small area estimation where there are many local areas with small (often zero) sample sizes from each area. The target is to estimate the local area means simultaneously, and very often a standard estimation procedure (which utilizes only the samples directly taken from a local area) can be improved by incorporating information from similar neighbouring areas.

To be specific, suppose the population is subdivided into strata $1, 2, \ldots, L$, where L is known. Attached to each unit i is a stratum membership h_i. In what follows, we will make a variety of assumptions about one's knowledge concerning $\boldsymbol{h} = (h_1, \ldots, h_N)^T$. For a given $\boldsymbol{h}$ we let $str_k(\boldsymbol{h}) = \{i : h_i = k\}$ be the units which belong to stratum k for $k = 1, \ldots, L$.

Also, let $N_k(h)$ denote the number of units in the population

which belong to stratum k. If s is a sample, let $str_k(\boldsymbol{h}(s))$ denote the number of units in the sample which belong to stratum k, and let $n_k(\boldsymbol{h}(s))$ denote the number of such units. Sometimes, when $\boldsymbol{h}$ is assumed to be known, we will just write str_k and N_k, and, if the sample is known as well, n_k. Our goal is to estimate simultaneously the stratum means $\mu_k = N_k^{-1} \sum_{i:h_i=k} y_i$.

The normal theory EB approach for solving this problem is as follows. Consider the model under which

(I) conditional on $\theta_1, \ldots, \theta_L, y_i$ are mutually independent, and for the units belonging to str_k, the y_i are normally distributed with mean θ_k and variance σ^2;

(II) $\theta_1, \ldots, \theta_L$ are iid $N(m, \tau^2)$.

Let $\bar{z}_k = n_k^{-1} \sum_{i \in str_k} y_i$, $\lambda = \sigma^2/\tau^2$, $B_k = \lambda/(\lambda + n_k)$. Also, let $\mathbf{1}_u$ denote a u component column vector with each element equal to 1. Then from Ericson (1969a) it follows that the joint posterior distribution of the y_i, $(i \in str_k,\ i \notin s)$, is $(N_k - n_k)$-variate normal with mean vector $((1 - B_k)\bar{z}_k + B_k m)\mathbf{1}_{N_k - n_k}$ and variance–covariance matrix $\sigma^2(\boldsymbol{I}_{N_k - n_k} + (\lambda + n_k)^{-1}\boldsymbol{J}_{N_k - n_k})$. Hence,

$$E(\mu_k|z) = N_k^{-1}[n_k\bar{z}_k + (N_k - n_k)((1 - B_k)\bar{z}_k + B_k m)]; \quad (4.4)$$

$$V(\mu_k|z) = N_k^{-1}(N_k - n_k)B_k(\tau^2 + \sigma^2/N_k). \quad (4.5)$$

In an EB analysis both λ and m are unknown, and are estimated from the marginal distribution of the y_i. Note that marginally the y_i are independent with the y_i belonging to str_k jointly distributed as $N(m\mathbf{1}_{N_k}, \sigma^2(\boldsymbol{I}_{N_k} + \lambda \boldsymbol{J}_{N_k}))$. Let $n_T = \sum_{k=1}^{L} n_k$ and $\bar{z} = \sum_{k=1}^{L} n_k\bar{z}_k/n_T$. At this point we require the assumption that $n_T > L$. Define

$$MSB = (L - 1)^{-1} \sum_{k=1}^{L} n_k(\bar{z}_k - \bar{z})^2; \quad (4.6)$$

$$MSW = (n_T - L)^{-1} \sum_{k=1}^{L} \sum_{i \in str_k} (z_i - \bar{z}_k)^2. \quad (4.7)$$

We shall see how the ratio MSB/MSW can be used in obtaining a consistent estimator of λ^{-1} if the n_k satisfy certain conditions.

With this end, first note that MSB and MSW are independently distributed. The following lemma gives the first two moments of MSB and MSW.

Lemma 4.3.1 *Consider the model given in (I) and (II). Then*

$$E(MSW) = \sigma^2; V(MSW) = 2\sigma^4(n_T - L)^{-1}; \tag{4.8}$$

$$E(MSB) = \sigma^2 + \tau^2(L-1)^{-1}\left(n_T - \sum n_k^2/n_T\right); \tag{4.9}$$

$$V(MSB) = 2\sigma^4(L-1)^{-2}\left[\sum_{k=1}^{L} B_k^{-2} + n_T^{-1}\left\{n_T^{-1}\left(\sum_{k=1}^{L} n_k B_k^{-1}\right)^2 - 2\sum_{k=1}^{L} n_k B_k^{-2}\right\}\right]. \tag{4.10}$$

Proof. First observe that $MSW \sim \sigma^2(n_T - L)^{-1}\chi^2_{n_T - L}$. This immediately leads to (4.8).

Next writing $W_k = n_k^{1/2}(\bar{z}_k - m)$ for $k = 1, \ldots, L$, it follows that $\boldsymbol{W} = (W_1, \ldots, W_L)^T \sim N(\boldsymbol{0}, \boldsymbol{D})$, where

$$\boldsymbol{D} = Diag(B_1^{-1}, \ldots, B_L^{-1}).$$

Now MSB can be expressed as

$$MSB = (L-1)^{-1}\boldsymbol{W}^T\boldsymbol{A}\boldsymbol{W},$$

where $\boldsymbol{A} = \boldsymbol{I}_L - \boldsymbol{g}\boldsymbol{g}^T, \boldsymbol{g}^T = n_T^{-1/2}(n_1^{1/2}, \ldots, n_L^{1/2})$. Using Corollary 1.1 of Searle (1971), it follows that

$$\begin{aligned} E[\boldsymbol{W}^T\boldsymbol{A}\boldsymbol{W}] &= \sigma^2 tr(\boldsymbol{A}\boldsymbol{D}) = \sigma^2[tr(\boldsymbol{D}) - \boldsymbol{g}^T\boldsymbol{D}\boldsymbol{g}] \\ &= \sigma^2[B_k^{-1} - \sum n_k B_k^{-1}/n_T] \\ &= \sigma^2[\sum(1 - n_k n_T^{-1})(\lambda + n_k)/\lambda] \\ &= \sigma^2(L-1) + \tau^2(n_T - n_T^{-1}\sum n_k^2). \end{aligned} \tag{4.11}$$

This leads to (4.9).

Next using Corollary 1.2 of Searle (1971), it follows that

$$V(\boldsymbol{W}^T\boldsymbol{A}\boldsymbol{W}) = 2\sigma^4(L-1)^{-2}[tr(\boldsymbol{A}\boldsymbol{D})^2]. \tag{4.12}$$

But

$$(\boldsymbol{A}\boldsymbol{D})^2 = (\boldsymbol{D} - \boldsymbol{g}\boldsymbol{g}^T\boldsymbol{D})^2 = \boldsymbol{D}^2 + \boldsymbol{g}\boldsymbol{g}^T\boldsymbol{D}\boldsymbol{g}\boldsymbol{g}^T\boldsymbol{D} - \boldsymbol{D}\boldsymbol{g}\boldsymbol{g}^T\boldsymbol{D} - \boldsymbol{g}\boldsymbol{g}^T\boldsymbol{D}^2, \tag{4.13}$$

so that

$$tr(\boldsymbol{A}\boldsymbol{D})^2 = tr(\boldsymbol{D}^2) + (\boldsymbol{g}^T\boldsymbol{D}\boldsymbol{g})^2 - 2\boldsymbol{g}^T\boldsymbol{D}^2\boldsymbol{g}$$

$$= \sum B_k^{-2} + n_T^{-1}[n_T^{-1}(\sum n_k B_k^{-1})^2 - 2\sum n_k B_k^{-2}]. \tag{4.14}$$

Combining (4.12) and (4.14), one gets (4.10). □

Using (4.10), it follows that assuming (a) $inf_{k\geq 1} n_k \geq 2$, one gets $V(MSW) \to 0$ as $L \to \infty$. Hence, MSW is an unbiased and consistent estimator of σ^2. Again, if one assumes (a) and (b) $sup_{k\geq 1} n_k < \infty$, then one gets $V(MSB) = O(L^{-1})$. Hence, a consistent estimator of λ^{-1} is given by

$$\hat{\lambda}^{-1} = max(0, (MSB/MSW - 1)(L-1)h^{-1}), \tag{4.15}$$

where $h = n_T - \sum n_k^2/n_T$. We modify the preceding estimator slightly so that the resulting estimator is identical with the James–Stein estimator in the balanced case, i.e. when $n_1 = \cdots = n_L$. Thus we propose the estimator

$$\hat{\lambda}^{-1} = max\left(0, [(L-1)MSB/((L-3)MSW) - 1]\,(L-1)h^{-1}\right) \tag{4.16}$$

for λ^{-1} for $L \geq 4$. Note that $\hat{\lambda}^{-1} \to \lambda^{-1}$ in probability as $L \to \infty$. Hence, $B_k = \lambda/(\lambda + n_k) = (1 + n_k\lambda^{-1})^{-1}$ is consistently estimated by $\hat{B}_k = (1 + n_k\hat{\lambda}^{-1})^{-1}, 1 \leq k \leq L$.

For later purposes, it will be important to prove the following lemma.

Lemma 4.3.2 *Under the assumptions (a) $inf_{k\geq 1} n_k \geq 2$, and (b) $sup_{k\geq 1} n_k = C < \infty$,*

$$max_{1\leq k\leq L}|\hat{B}_k - B_k| \to 0 \text{ in probability as } L \to \infty. \tag{4.17}$$

Proof. $|\hat{B}_k - B_k| = n_k|\hat{\lambda}^{-1} - \lambda^{-1}|/[(1 + n_k\hat{\lambda}^{-1})(1 + n_k\lambda^{-1})] \leq C|\hat{\lambda}^{-1} - \lambda^{-1}|$. Since the right-side of this inequality is the same for each $k = 1, \ldots, L$, it follows that $max_{1\leq k\leq L}|\hat{B}_k - B_k| \leq C|\hat{\lambda}^{-1} - \lambda^{-1}|$. Since $\hat{\lambda}^{-1} \to \lambda^{-1}$ in probability as $L \to \infty$, the lemma follows. □

Next to estimate m based on the marginal distribution of the y_i, first assume that λ is known.Then the MLE of m is given by

$$\begin{aligned} \tilde{m} &= \sum n_k \bar{z}_k(\sigma^2 + \tau^2 n_k)^{-1} / \sum n_k(\sigma^2 + \tau^2 n_k)^{-1} \\ &= \sum n_k(\lambda + n_k)^{-1}\bar{z}_k / \sum n_k(\lambda + n_k)^{-1} \\ &= \sum (1 - B_k)\bar{z}_k / \sum (1 - B_k). \end{aligned} \tag{4.18}$$

Note that the condition $\hat{M}^{-1} = 0$ is equivalent to the condition that $\hat{B}_1 = \cdots = \hat{B}_L = 1$. Thus, motivated from (4.18), we propose the EB estimator

$$\hat{m} = \begin{cases} \sum(1 - \hat{B}_k)\bar{z}_k / \sum(1 - \hat{B}_k) & \text{if } \hat{\lambda}^{-1} \neq 0 \\ L^{-1} \sum \bar{z}_k & \text{if } \hat{\lambda}^{-1} = 0. \end{cases} \tag{4.19}$$

for m. Substituting the estimators $\hat{m}$ and $\hat{B}_k$ respectively for m and B_k, it follows that an EB predictor of $\mu_k = N_k^{-1} \sum_{i \in str_k} y_i$ is given by

$$\hat{\mu}_k^{EB} = N_k^{-1}[n_k \bar{z}_k + (N_k - n_k)((1 - \hat{B}_k)\bar{z}_k + \hat{B}_k \hat{m})]. \tag{4.20}$$

It is conceivable to think of procedures alternative to the proposed method of EB estimation. A natural candidate is an EB estimator based on MLE's of λ^{-1} and m. It can be shown (see e.g. Herbach, 1959) that when $n_1 = \cdots = n_L = n$, the MLE of m is $\bar{z}$ which is identical to $\hat{m}$. In this case, the MLE of λ^{-1} is given by

$$\hat{\hat{\lambda}}^{-1} = max(0, ((L-1)L^{-1}MSB/MSW - 1)n^{-1}). \tag{4.21}$$

The estimator $\hat{\lambda}^{-1}$ simplifies in this case to $\hat{\lambda}^{-1} = max(0, ((L - 1)MSB/((L - 3)MSW) - 1)n^{-1})$ since $h = n(L - 1)$ in this case. The estimators $\hat{\lambda}^{-1}$ and $\hat{\hat{\lambda}}^{-1}$ differ only in the multipliers of MSB/MSW, and clearly they have similar asymptotic $L \to \infty$ performance. In the unbalanced case (i.e. when the n_k's are not all equal) the ML approach is analytically intractable, since the likelihood equations do not admit any closed-form solutions.

Next we compare the Bayes risk (integrated risk over both the samples and the parameters) performance of the EB estimator $\hat{\boldsymbol{\mu}}^{EB} = (\hat{\mu}_1^{EB}, \ldots, \hat{\mu}_L^{EB})^T$ of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^T$ with the classical estimator $\hat{\boldsymbol{\mu}}_0 = (\bar{z}_1, \ldots, \bar{z}_L)^T$ Denote the prior distribution given in (I) and (II) by ξ. We shall consider the average squared error loss

$$L(\boldsymbol{\mu}, \boldsymbol{e}) = L^{-1} \sum_{k=1}^{L} (\mu_k - e_k)^2 \tag{4.22}$$

for an arbitrary estimator $\boldsymbol{e} = (e_1, \ldots, e_L)^T$ of $\boldsymbol{\mu}$. Let $r(\xi, \boldsymbol{e})$ denote the Bayes risk of $\boldsymbol{e}$. Rather than comparing $r(\xi, \hat{\boldsymbol{\mu}}^{EB})$ directly with $r(\xi, \hat{\boldsymbol{\mu}}_0)$, we provide a normalized version of the Bayes risk improvement of $\hat{\boldsymbol{\mu}}^{EB}$ over $\hat{\boldsymbol{\mu}}_0$.

With this end, we introduce the following notion of **relative savings loss** (RSL) as given in Efron and Morris (1973). The

RSL of $\hat{\boldsymbol{\mu}}^{EB}$ with respect to an arbitrary estimator $\boldsymbol{e}$ of $\boldsymbol{\mu}$ is given by

$$RSL(\xi; \hat{\boldsymbol{\mu}}^{EB}, \boldsymbol{e}) = [r(\xi, \hat{\boldsymbol{\mu}}^{EB}) - r(\xi, \hat{\boldsymbol{\mu}}^{B})]/[r(\xi, \boldsymbol{e}) - r(\xi, \hat{\boldsymbol{\mu}}^{B})]. \tag{4.23}$$

where $\hat{\boldsymbol{\mu}}^{B}$ denotes the Bayes estimator of $\boldsymbol{\mu}$ under the prior given in (I) and (II), and the loss given in (4.22). Thus $RSL(\xi; \hat{\boldsymbol{\mu}}^{EB}, \boldsymbol{e})$ denotes the proportion of possible Bayes risk improvement over $\boldsymbol{e}$ that is sacrificed by the use of the EB estimator $\hat{\boldsymbol{\mu}}^{EB}$ instead of the Bayes estimator $\hat{\boldsymbol{\mu}}^{B}$ when the prior is ξ. [Of course, if $r(\xi, \boldsymbol{e}) < r(\xi, \hat{\boldsymbol{\mu}}^{B})$, then $\boldsymbol{e}$ is superior to $\hat{\boldsymbol{\mu}}^{EB}$. The RSL expression given in (4.23) is meaningful only when $r(\xi, \hat{\boldsymbol{\mu}}^{EB}) < r(\xi, \boldsymbol{e})$.] Using the well-known result

$$r(\xi, \boldsymbol{e}) = r(\xi, \hat{\boldsymbol{\mu}}^{B}) + E||\boldsymbol{e} - \hat{\boldsymbol{\mu}}^{B}||^2 \tag{4.24}$$

one now gets the following theorem.

Theorem 4.1 *Under the prior ξ given in (I) and (II),*

$$\begin{aligned} RSL(\xi; \hat{\boldsymbol{\mu}}^{EB}, \hat{\boldsymbol{\mu}}_0) &= \sum_{k=1}^{L} E[(\hat{B}_k - B_k)(\bar{z}_k - m) \\ &- \hat{B}_k(\hat{m} - m)]^2 / \sum_{k=1}^{L} B_k^2 E(\bar{z}_k - m)^2. \end{aligned} \tag{4.25}$$

Next we examine the asymptotic (as $L \to \infty$) behaviour of the RSL expression given in (4.25). The following theorem is proved.

Theorem 4.2 *Under the assumptions (a) $inf_{k \geq 1}\, n_k \geq 2$ and (b) $sup_{k \geq 1} n_k = C < \infty$,*

$$RSL(\xi; \hat{\boldsymbol{\mu}}^{EB}, \hat{\boldsymbol{\mu}}_0) \to 0 \text{ as } L \to \infty. \tag{4.26}$$

Proof. In view of the assumptions of the theorem, it follows that

$$\sum B_k^2 E[\bar{z}_k - m]^2 \geq L\lambda^2(\lambda + C)^{-2}[\tau^2 + \sigma^2/C]. \tag{4.27}$$

Hence, in view of (4.25), we prove the theorem by showing that

$$L^{-1} \sum_{k=1}^{L} E[(\hat{B}_k - B_k)(\bar{z}_k - m)]^2 \to 0 \text{ as } L \to \infty; \tag{4.28}$$

$$L^{-1}\sum E[\hat{B}_k(\hat{m}-m)]^2 \to 0 \text{ as } L \to \infty. \tag{4.29}$$

We first prove (4.28). Note that

$$\sum(\hat{B}_k - B_k)^2(\bar{z}_k - m)^2 \leq max_{1\leq k\leq L}(\hat{B}_k - B_k)^2 L^{-1}\sum(\bar{z}_k - m)^2. \tag{4.30}$$

Note now that L^{-1} times the right-hand side of (4.30) converges to zero in probability as $L \to \infty$ in view of Lemma 4.3.2, and the fact that

$$\begin{aligned} L^{-1}\sum E(\bar{z}_k - m)^2 &= L^{-1}\sum(\tau^2 + \sigma^2/n_k) \\ &\leq (\tau^2 + \sigma^2). \end{aligned} \tag{4.31}$$

Also, using the Schwarz inequality,

$$E[L^{-1}\sum(\bar{z}_k - m)^2]^2 \leq L^{-1}\sum E(\bar{z}_k - m)^4 = O(1). \tag{4.32}$$

Using (4.30), Lemma 4.3.2, the fact that $max_{1\leq k\leq L}|\hat{B}_k - B_k| \leq 1$, and (4.32), it follows that the left-hand side of (4.28) converges to zero in probability as $L \to \infty$, and it is uniformly integrable in $L \geq 1$. This proves (4.28).

Next, to prove (4.29), first use the inequality

$$L^{-1}\sum \hat{B}_k{}^2(\hat{m}-m)^2 \leq (\hat{m}-m)^2. \tag{4.33}$$

Now observe that using assumption (a) of the theorem, $1 - \hat{B}_k = n_k/(\hat{\lambda} + n_k) \geq 2/(\hat{\lambda} + 2)$. Hence,

$$\begin{aligned} |\hat{m} - m| &= |\sum(1 - \hat{B}_k)(\bar{z}_k - m)|/\sum(1 - \hat{B}_k) \\ &\leq (1/2)(\hat{\lambda} + 2)L^{-1}|\sum(1 - \hat{B}_k)(\bar{z}_k - m)|. \end{aligned} \tag{4.34}$$

But

$$\begin{aligned} |\sum(1 - \hat{B}_k)(\bar{z}_k - m)| &\leq |\sum(1 - B_k)(\bar{z}_k - m)| \\ &+ max_{1\leq k\leq L}|\hat{B}_k - B_k|\sum|\bar{z}_k - m|. \end{aligned} \tag{4.35}$$

Note that using Lemma 4.3.2, $max_{1\leq k\leq L}|\hat{B}_k - B_k|$ converges to zero in probability as $L \to \infty$. Also, $L^{-1}\sum|\bar{z}_k - m|$ is bounded in probability. Finally,

$$E[L^{-1}\sum(1 - B_k)(\bar{z}_k - m)]^2 = L^{-2}\sum(1 - B_k)^2(\tau^2 + \sigma^2/n_k)$$

$$\leq \quad L^{-1}(\tau^2 + \sigma^2) \to 0 \qquad (4.36)$$

as $L \to \infty$. This implies that $L^{-1}\sum(1 - B_k)(\bar{z}_k - m)$ converges to zero in probability as $L \to \infty$. Hence,

$$\hat{m} - m \ \to 0 \text{ in probability as } L \to \infty. \qquad (4.37)$$

Moreover, using assumptions (I) and (II) of the theorem, and the Schwarz inequality,

$$\begin{aligned} |\hat{m} - m|^2 &\leq (C/(\hat{\lambda} + C))^2(\hat{\lambda} + 2)^2/4]L^{-1}\sum|\bar{z}_k - m|^2 \\ &\leq (1/4)max(1, C^2)L^{-1}\sum|\bar{z}_k - m|^2 \qquad (4.38) \end{aligned}$$

which proves the uniform integrability of $(\hat{m} - m)^2$. Combining (4.33), (4.37) and (4.38) one gets (4.29)and the proof is complete. □

Remark 1. Since $L^{-1}\sum B_k^2 E(\bar{z}_k - m)^2 \geq (M/(M + C))^2\tau^2$, it follows from (4.2) that $r(\xi, \hat{\mu}^{EB}) - r(\xi, \hat{\mu}^B) \to 0$ as $L \to \infty$. Thus the proposed EB procedure is **asymptotically optimal** in the sense of Robbins (1956) .

Remark 2. In the special case when $n_1 = \cdots = n_L$ it can be shown that $RSL(\xi; \hat{\mu}^{EB}, \hat{\mu}_0) = O(L^{-1})$. The details are omitted.

Remark 3. In the case when M (and, hence, the B_k's) are known, and m is the only unknown parameter, one estimates m by $\tilde{m}$ as given in (4.18). It is now easy to see that

$$RSL(\xi; \hat{\mu}^{EB}, \hat{\mu}_0) = \sum B_k^2 E(\hat{m} - m)^2/[\sum B_k^2 E(\bar{z}_k - m)^2]. \quad (4.39)$$

It is easy to verify in this case that $RSL(\xi; \hat{\mu}^{EB}, \hat{\mu}_0) = O(L^{-1})$. In the further special case when $n_1 = \cdots = n_L$, this RSL expression is exactly equal to L^{-1}.

Certain generalizations of the above model seem feasible. For instance, Little (1983) has considered the analysis of disproportionate stratified samples from a model-based Bayesian perspective. The major difference between Little's approach with the one discussed in this section is that rather than using the same σ^2 across all the strata, he assigns different first-stage variances for different strata. These first-stage variances are generated from a diffuse inverse gamma prior. The inclusion of distinct stratum variances overcomes distortions in the sample introduced by different selection probabilities. In particular, the model-based estimators become **design-consistent** estimators.

In the next section, we shall dispense with the normality assumption in the Bayesian model, and study robustness of the proposed EB procedure.

4.4 Robust estimation of stratum means

We considered in the previous section EB estimation of the finite population mean assuming a normal superpopulation model. In this section, the normality assumption is relaxed, and is replaced instead by the assumption of posterior linearity as introduced in (1.5). As in the previous section, the population is subdivided into strata $1, \ldots, L$, where L is known. Also, the notations introduced at the beginning of the previous section to identify a stratum or a unit within a stratum still hold. As before, we want to estimate the vector of stratum means under the loss given in (4.22).

The model that we are going to assume is as follows:

(I) Conditional on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^T$, y_i are mutually independent, and for the units belonging to str_k, $E(y_i|\boldsymbol{\theta}) = \theta_k$ and $V(y_i) = \mu_2(\theta_k)$, $k = 1, \ldots, L$;

(II) θ_k are iid with mean m and variance τ^2;

(III) $0 < \sigma^2 = E(\mu_2(\theta_k)) < \infty$.

The basic assumption of this section is that

$$E(\theta_k|z) = \sum_{i \in str_k} a_{ki} z_i + b_k, \; k = 1, \ldots, L, \tag{4.40}$$

where the a_{ki} and the b_k are constants not depending on the y_i. In view of the fact that for $i \in str_k$, the y_i are iid given $\boldsymbol{\theta}$, from Goldstein (1975), (4.40) leads to

$$E(\theta_k|z) = a_k \bar{z}_k + b_k, \; k = 1, \ldots, L, \tag{4.41}$$

where the a_k are constants. Henceforth, we shall assume (4.41) instead of (4.40).

Several comments about this model seem to be in order. First, if conditional on the θ_k, the y_i belong to the one-parameter exponential family, and the posterior linearity assumption as given in (4.41) holds, then it follows from Diaconis and Ylvisaker (1979) that the θ_k belong to the conjugate prior family. Second, if conditional on the θ_k, the y_i belong to the **natural exponential family** (see Morris, 1982, 1983c), then the $\mu_2(\theta_k)$ determine the appropriate family of distributions of the θ_k.

We now turn to the Bayes estimation of stratum means under the model given in (I)–(III), and the posterior linearity assumption as given in (4.41). Using the notations $\lambda = \sigma^2/\tau^2$ and $B_k = \lambda/(\lambda+n_k)$ as in the previous section, it follows from Goldstein (1975) that $a_k = \tau^2/(\tau^2 + \sigma^2 n_k^{-1}) = n_k/(\lambda + n_k) = 1 - B_k$, and $b_k = B_k m$, $k = 1, \ldots, L$. Hence, under the loss given in (4.22), the Bayes estimator of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^T$ is $\hat{\boldsymbol{\mu}}^B = (\hat{\mu}_1^B, \ldots, \hat{\mu}_L^B)^T$, where

$$\begin{aligned}
\hat{\mu}_k^B &= E(\mu_k|z) \\
&= N_k^{-1}\left[n_k\bar{z}_k + \sum_{i\in str_k, i\notin s} E(y_i|z)\right] \\
&= N_k^{-1}\left[n_k\bar{z}_k + \sum_{i\in str_k, i\notin s} E((y_i|\boldsymbol{\theta}, z)|z)\right] \\
&= N_k^{-1}[n_k\bar{z}_k + (N_k - n_k)E(\theta_k|z)] \\
&= N_k^{-1}[n_k\bar{z}_k + (N_k - n_k)((1 - B_k)\bar{z}_k + B_k m)] \\
&= (1 - f_k B_k)\bar{z}_k + f_k B_k m \\
&= \bar{z}_k - f_k B_k(\bar{z}_k - m), \qquad (4.42)
\end{aligned}$$

where $f_k = (N_k - n_k)/N_k$ denotes the finite population correction for stratum k.

It is pointed out in Ericson (1969b) that (4.41) holds for many parametric families of priors including but not limited to the normal prior. It is also shown in Goldstein (1975) that (4.41) also holds for Dirichlet process priors (see Ferguson, 1973). Yet another example of a parametric family of priors where (4.41) holds is when

(i) conditional on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^T$ and h, for units belonging to stratum k, y_i are iid $N(\theta_k, h^{-1})$;

(ii) conditional on h, θ_k are iid $N(m, h^{-1})$;

(iii) h has a proper prior distribution.

In particular, when h has a gamma distribution, then the joint distribution of the y_i belonging to a particular stratum is multivariate-t. Unlike our earlier assumption, in this case the θ_k are only iid (normal) conditional on h. But the results of this section based on the iid assumption of the θ_k continue to hold by using the simple device of first conditioning on h.

In an EB analysis, σ^2, τ^2 and m are all unknown, and need to be estimated from the data. However, first we consider the simple

case when σ^2 and τ^2 (and hence λ) is known, but m is unknown, and needs to be estimated. Note that marginally

$$E(\bar{z}_k) = EE(\bar{z}_k|\boldsymbol{\theta}) = E(\theta_k) = m; \tag{4.43}$$

$$\begin{aligned} V(\bar{z}_k) &= E[V(\bar{z}_k)|\boldsymbol{\theta}] + V[E(\bar{z}_k)|\boldsymbol{\theta}] \\ &= E[\mu_2(\theta_k)n_k^{-1}] + V(\theta_k) \\ &= \sigma^2 n_k^{-1} + \tau^2 = \sigma^2\lambda^{-1}(1 - B_k)^{-1}. \end{aligned} \tag{4.44}$$

Hence, the BLUE of m is given by

$$\tilde{m} = \sum_{k=1}^{L}(1 - B_k)\bar{z}_k / \sum_{k=1}^{L}(1 - B_k). \tag{4.45}$$

Note that if the underlying distribution is normal, the BLUE is identical to the MLE of M. The resulting EB estimator of $\boldsymbol{\mu}$ is $\tilde{\boldsymbol{\mu}}^{EB} = (\tilde{\mu}_1^{EB}, \ldots, \tilde{\mu}_L^{EB})^T$, where

$$\tilde{\mu}_k^{EB} = \bar{z}_k - f_k B_k(\bar{z}_k - \tilde{m}),\ k = 1, \ldots, L. \tag{4.46}$$

It is interesting to note that if $n_1 = \cdots = n_L = n$, then $B_1 = \cdots = B_L = B = \lambda/(\lambda + n)$, and $\tilde{m} = L^{-1}\sum_{k=1}^{L}\bar{z}_k = \bar{z}$ (say). Then

$$\tilde{\mu}_k^{EB} = \bar{z}_k - f_k B(\bar{z}_k - \bar{z}),\ k = 1, \ldots, L \tag{4.47}$$

which is the finite population analogue of an estimator of Lindley and Smith (1972) that was motivated under the normality assumption using an HB approach.

Next we compare the performance of $\tilde{\boldsymbol{\mu}}^{EB}$ with the classical estimator $\tilde{\boldsymbol{\mu}}_0 = (\bar{z}_1, \ldots, \bar{z}_L)^T$ of $\boldsymbol{\mu}$ in terms of their Bayes risks. Denote the prior given in (I)–(III) by ξ. As in the previous section, let $r(\xi, \boldsymbol{e})$ denote the Bayes risk of an estimator $\boldsymbol{e}$ of $\boldsymbol{\mu}$ with respect to the prior ξ. Recall the definition of **relative savings loss** as given in the previous section. Using the facts that under the loss given in (4.22), $r(\xi, \boldsymbol{e}) = r(\xi, \tilde{\boldsymbol{\mu}}^B) + L^{-1}E||\boldsymbol{e} - \tilde{\boldsymbol{\mu}}^B||^2$ and $V(\bar{z}_k) = \sigma^2\lambda^{-1}(1 - B_k)^{-1}$, one gets

$$\begin{aligned} &RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0) \\ &\quad = \lambda^{-1}(\sum f_k^2 B_k^2)(\sum(1 - B_k))^{-1}(\sum f_k^2 B_k n_k^{-1})^{-1}. \end{aligned} \tag{4.48}$$

Our next theorem shows that as $L \to \infty$, under very mild conditions, $RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0) \to 0$.

Theorem 4.3 *Under the assumptions that (a)* $inf_{k\geq 1} n_k \geq 1$, *and (b)* $sup_{k\geq 1} n_k = C < \infty$, $RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0) = O(L^{-1})$ *as* $L \to \infty$.

Proof. Use $\sum_{k=1}^{L} f_k^2 B_k^2 \leq L$. Also, using assumption (a) of the theorem, $\sum_{k=1}^{L}(1-B_k) = \sum_{k=1}^{L} n_k/(\lambda+n_k) \geq L/(\lambda+1)$. Further, using both (a) and (b),

$$\begin{aligned}
\sum_{k=1}^{L} f_k^2 B_k^2 n_k^{-1} &= \sum_{k=1}^{L} N_k^{-2}(N_k - n_k)^2 n_k(\lambda + n_k)^{-2} \\
&\geq \sum_{k=1}^{L}(1 - n_k(n_k+1)^{-1})^2 n_k(\lambda + n_k)^{-2} \\
&\geq (1 - C(C+1)^{-1})^2 \sum_{k=1}^{L}(\lambda + n_k)^{-2} \\
&\geq (1 - C(C+1)^{-1})^2 L(\lambda + C)^{-2}.
\end{aligned}$$

Now, from (4.48), one gets $RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0) \leq \lambda^{-1} L(\lambda+1)^{-1} L^{-1}$ $(1 - C(C+1)^{-1})^{-2} L^{-1}(\lambda + C)^2 = O(L^{-1})$. □

In the course of proving the theorem, we have also shown that $r(\xi, \tilde{\boldsymbol{\mu}}_0) - r(\xi, \tilde{\boldsymbol{\mu}}^B) = O(1)$, and $r(\xi, \tilde{\boldsymbol{\mu}}^{EB}) - r(\xi, \tilde{\boldsymbol{\mu}}^B) \to 0$ as $L \to \infty$. Thus, for known σ^2 and τ^2, the proposed EB procedure is **asymptotically optimal** in the sense of Robbins (1956). Also, in the special case when $n_1 = \cdots = n_L = n$, so that $B_1 = \cdots = B_L = B$, $RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0) = L^{-1}$ for all L.

Next we consider the more realistic situation when m, σ^2, and τ^2 are all unknown, and need to be estimated. As in the previous section, write $n_T = \sum_{k=1}^{L} n_k$. It is assumed that $n_T > L$. Also, recall the notations $\bar{z} = \sum n_k \bar{z}_k / n_T$, $SSB = \sum n_k(\bar{z}_k - \bar{z})^2$, $MSB = SSB/(L-1)$, $SSW = \sum_{k=1}^{L} \sum_{i \in str_k, i \in s}(y_i - \bar{z}_k)^2$, $MSW = SSW/(n_T - L)$. We now prove a result which motivates estimation of λ^{-1}. The result was proved in the previous section under the normality assumption. However, in this section, the result is proved only under the model given in (I)–(III), without bringing in any distributional assumptions.

Lemma 4.4.1 *Consider the model given in (I)–(III). Then*

$$E(MSW) = \sigma^2;\ E(MSB) = \sigma^2 + g\tau^2(L-1)^{-1}, \tag{4.49}$$

where $g = n_T - \sum n_k^2 / n_T$.

The proof of the above lemma, and many of the subsequent results, hinge crucially on the following lemma, which is proved below.

Lemma 4.4.2 *Let* $W_1, \ldots, W_L$ *be independent random variables, with* $E(W_k) = 0$, $E(W_k^2) = \mu_{2k}$, $E(W_k^4) = \mu_{4k}$, *where* $0 < \mu_{2k} <$

$\mu_{4k}^{1/2} < \infty$, $k = 1, \ldots, L$. *Write* $\boldsymbol{W} = (W_1, \ldots, W_L)^T$. *Then for any symmetric* $L \times L$ *matrix* $\boldsymbol{A} = ((a_{kk'}))$, *one has*

$$E(\boldsymbol{W}^T \boldsymbol{A} \boldsymbol{W}) = \sum_{k=1}^{L} a_{kk} \mu_{2k}; \tag{4.50}$$

$$\begin{aligned} V(\boldsymbol{W}^T \boldsymbol{A} \boldsymbol{W}) &= \sum_{k=1}^{L} a_{kk}^2 (\mu_{4k} - \mu_{2k}^2) \\ &+ 2 \sum_{1 \le k \ne k' \le L} \sum a_{kk'}^2 \mu_{2k} \mu_{2k'} E(W_k W_{k'}). \end{aligned} \tag{4.51}$$

Proof. First using the independence of the W_k and the fact that $E(W_k) = 0$ one gets

$$\begin{aligned} E(\boldsymbol{W}^T \boldsymbol{A} \boldsymbol{W}) &= \sum_{k=1}^{L} a_{kk} E(W_k^2) + \sum_{1 \le k \ne k' \le L} \sum a_{kk'} E(W_k W_{k'}) \\ &= \sum_{k=1}^{L} a_{kk} \mu_{2k}; \end{aligned}$$

$$\begin{aligned} V(\boldsymbol{W}^T \boldsymbol{A} \boldsymbol{W}) &= V\left(\sum_{k=1}^{L} a_{kk} W_k^2 + \sum_{1 \le k \ne k' \le L} \sum a_{kk'} W_k W_{k'} \right) \\ &= V\left(\sum_{k=1}^{L} a_{kk} W_k^2 \right) + 2 \sum_{1 \le k \ne k' \le L} \sum a_{kk'}^2 E(W_k^2) E(W_{k'}^2) \\ &= \sum_{k=1}^{L} a_{kk}^2 (\mu_{4k} - \mu_{2k}^2) + 2 \sum_{1 \le k \ne k' \le L} \sum a_{kk'}^2 \mu_{2k} \mu_{2k'}. \end{aligned}$$

□

We now prove Lemma 4.4.1.

Proof. Write $\boldsymbol{z}_k$ as the vector of y_i such that $i \in str_k$, and $i \in s$. Then it is possible to re-express SSW as

$$SSW = \sum_{k=1}^{L} \boldsymbol{z}_k^T (\boldsymbol{I}_{n_k} - n_k^{-1} \boldsymbol{J}_{n_k}) \boldsymbol{z}_k.$$

Since conditional on $\boldsymbol{\theta}$, the components of $\boldsymbol{z}_k$ have a common mean

θ_k and a common variance $\mu_2(\theta_k)$, it follows that

$$E(MSW) = \sum_{k=1}^{L}(n_k - 1)E(\mu_2(\theta_k)) = (L-1)\sigma^2.$$

This proves the first part of the lemma. To prove the second part, we write

$$E(SSB) = \sum n_k E[(\bar{z}_k - \bar{z})^2 + (\theta_k - \bar{\theta})^2 + 2(\bar{z}_k - \bar{z})(\theta_k - \bar{\theta})],$$

where $\bar{\theta} = \sum_{k=1}^{L} n_k\theta_k/n_T$. Note that $E(\bar{z}_k - \bar{z}|\boldsymbol{\theta}) = 0$. Further, applying Lemma 4.4.2 with $\boldsymbol{W} = (\bar{z}_1, \ldots, \bar{z}_L)^T$ and

$$\boldsymbol{A} = diag(n_1, \ldots, n_L) - n_T^{-1}\boldsymbol{n}\boldsymbol{n}^T,$$

where $\boldsymbol{n}^T = (n_1, \ldots, n_L)$, it follows after some algebra that

$$E[\sum n_k(\bar{z}_k - \bar{z})^2|\boldsymbol{\theta}] = \sum(1 - n_k n_T^{-1})\mu_2(\theta_k).$$

Using $E(\mu_2(\theta_k)) = \sigma^2$, one now gets

$$E\left[\sum_{k=1}^{L} n_k(\bar{z}_k - \bar{z})^2\right] = (L-1)\sigma^2.$$

Similarly, applying Lemma 4.4.2 to $\theta_k - \mu$'s, one gets

$$E\left[\sum_{k=1}^{L} n_k(\theta_k - \bar{\theta})^2\right] = g\tau^2.$$

Thus,

$$E(SSB) = (L-1)\sigma^2 + g\tau^2.$$

□

It follows from Lemma 4.4.1 that a suitable estimator of σ^2 is MSW, while the method of moments estimator of τ^2 is

$$max(0, (MSB - MSW)(L-1)g^{-1}).$$

Accordingly, a sensible estimator of $\lambda^{-1} = \tau^2/\sigma^2$ is

$$\hat{\lambda}^{-1} = max(0, [(L-1)MSB/((L-3)MSW) - 1](L-1)g^{-1}). \tag{4.52}$$

The above estimator of λ^{-1} is identical to the one given in Ghosh and Meeden (1986). The corresponding estimator of B_k is $\hat{B}_k = (1 + n_k\hat{\lambda}^{-1})^{-1}$, $k = 1, \ldots, L$. The multiplier $(L-1)/(L-3)$ of MSB/MSW makes the resulting EB estimator identical to the

positive part James–Stein estimator (James and Stein, 1961) when $n_1 = \cdots = n_L = n$. Now, in view of (4.52), we estimate m by

$$\hat{m} = \begin{cases} \sum_{k=1}^{L}(1-\hat{B}_k)\hat{z}_k / \sum_{k=1}^{L}(1-\hat{B}_k) & \text{if } \hat{\lambda}^{-1} \neq 0 \\ \bar{z} & \text{otherwise.} \end{cases} \tag{4.53}$$

Hence, an EB estimator of μ is $\hat{\mu}^{EB}$ where

$$\hat{\mu}_k^{EB} = \bar{z}_k - f_k \hat{B}_k(\bar{z}_k - \hat{m}),\ k = 1, \ldots, L. \tag{4.54}$$

Under the loss given in (4.22), the Bayes risk of $\hat{\mu}^{EB}$ is given by

$$r(\xi, \hat{\mu}^{EB}) = r(\xi, \hat{\mu}^{B}) + L^{-1}\sum f_k^2 E[(\hat{B}_k - B_k)(\bar{z}_k - m) - \hat{B}_k(\hat{m} - m)]^2. \tag{4.55}$$

The next main result of this section is as follows.

Theorem 4.4 *Under the assumptions that (a) $inf_{k\geq 1} n_k \geq 2$, (b) $sup_{k\geq 1} n_k = C < \infty$, (c) $E[\mu_4(\theta_1)] < \infty$, where $\mu_4(\theta_k) = E[(y_i - \theta_k)^2 | \theta_k]$ for $i \in str_k$, and (d) $E(\theta_1^4) < \infty$,*

$$r(\xi, \hat{\mu}^{EB}) - r(\xi, \hat{\mu}^{B}) \to 0 \text{ as } L \to \infty. \tag{4.56}$$

Remark 1. Assumptions (c) and (d) hold automatically for normal priors, and were not needed explicitly in the previous section. In addition, in many instances, (d) implies (c). For example, in the natural exponential family with quadratic variance functions, $\mu_2(\theta_1)$ is at most a quadratic in θ_1 so that (d) implies (c) automatically. Also, it follows as an immediate corollary of Theorem 4.4 that $RSL(\xi; \hat{\mu}^{EB}, \hat{\mu}_0) \to 0$ as $L \to \infty$.

The proof of the theorem is greatly facilitated by the following lemma.

Lemma 4.4.3 *Under the assumptions of Theorem 4.4, $\hat{\lambda} \to \lambda$ in probability as $L \to \infty$.*

Proof. We prove the result by showing that (a) $MSW \to \sigma^2$ and (b) $MSB - E(MSB) \to 0$ in probability as $L \to \infty$. Then, in view of Lemma 4.4.1, it follows that

$$MSB/MSW - 1 - g(L-1)^{-1}\lambda^{-1} \to 0 \text{ in probability as } L \to \infty.$$

Hence, $(L-1)g^{-1}[MSB/MSW - 1] \to \lambda^{-1}$ in probability as $L \to$

∞. Further,

$$g = \sum\sum_{1 \le k \ne k' \le L} n_k n_{k'} / n_T \ge 4(L-1)^{-1} M^{-1} C^{-1}.$$

Hence, $[(L-1)MSB/((L-3)MSW) - 1](L-1)g^{-1} \to \lambda^{-1}$ in probability as $L \to \infty$.

It remains to prove (a) and (b). To prove (a), first we get an explicit expression for $V(SSW) = \sum_{k=1}^{L} V(\sum_{i \in str_k, i \in s}(y_i - \bar{z}_k)^2)$. Using Lemma 4.4.2 it follows after some algebra that

$$V\left[\sum_{i \in str_k, i \in s}(y_i - \bar{z}_k)^2 | \theta_k\right] = n_k^{-1}(n_k - 1)[\mu_4(\theta_k)(n_k - 1) - \mu_2^2(\theta_k)(n_k - 3)]. \quad (4.57)$$

Now using

$$E\left[\sum_{i \in str_k, i \in s}(y_i - \bar{z}_k)^2 | \theta_k\right] = (n_k - 1)\mu_2(\theta_k)$$

and

$$V[\mu_2(\theta_k)] = E[\mu_2^2(\theta_1)] - \sigma^4,$$

one gets

$$V\left[E\left(\sum_{i \in str_k, i \in s}(y_i - \bar{z}_k)^2 | \theta_k\right)\right] = (n_k - 1)^2[E\mu_2^2(\theta_1) - \sigma^4]. \quad (4.58)$$

Now combining (4.57) and (4.58), one gets after some simplifications,

$$\begin{aligned} V(SSW) &= \left[\sum_{k=1}^{L}(n_k - 1)^2 n_k^{-1}\right] E\mu_4(\theta_1) \\ &+ \left[\sum_{k=1}^{L}\{(n_k - 1)^2 - (n_k - 1)(n_k - 3)n_k^{-1}\}\right] E\mu_2^2(\theta_1) \\ &- \left[\sum_{k=1}^{L}(n_k - 1)^2\right]\sigma^4. \end{aligned} \quad (4.59)$$

Hence, under assumptions (a) and (b), $V(SSW) \le CL$, where in the above, and in what follows, C is a generic constant. Now, using

$n_T \geq 2L$ (from assumption (a)), one gets

$$V(MSW) \leq CL(n_T - L)^{-2} \leq CL^{-1}, \tag{4.60}$$

which proves that $MSW - E(MSW) \to 0$ in probability as $L \to \infty$, that is $MSW \to \sigma^2$ in probability as $L \to \infty$.

It remains only to show that $V(MSB) \to 0$ as $L \to \infty$. Using the inequality $V(SSB) \leq 3[V(U_1) + V(U_2) + V(U_3)]$, and using the representation of SSW as in the proof of Lemma 4.4.1, it follows that

$$\begin{aligned} V(SSB) &\leq 3\left[V\left\{\sum_{k=1}^{L} n_k(\bar{z}_k - \bar{z})^2\right\}\right. \\ &+ V\left\{\sum_{k=1}^{L} n_k(\theta_k - \bar{\theta})^2\right\} \\ &+ \left. 4V\left\{\sum_{k=1}^{L} n_k(\bar{z}_k - \bar{z})(\theta_k - \bar{\theta})\right\}\right]. \end{aligned} \tag{4.61}$$

Next, using $E(\bar{z}_k|\theta_k) = 0$, one gets

$$\begin{aligned} & V\left[\sum_{k=1}^{L} n_k(\bar{z}_k - \bar{z})(\theta_k - \bar{\theta})\right] \\ &= V[n_k\bar{z}_k(\theta_k - \bar{\theta})] \\ &= E\left[V\left\{\sum_{k=1}^{L} n_k\bar{z}_k(\theta_k - \bar{\theta})|\boldsymbol{\theta}\right\}\right] \\ &= E\left[\sum_{k=1}^{L} n_k\mu_2(\theta_k)(\theta_k - \bar{\theta})^2\right] \\ &\leq \sum_{k=1}^{L} n_k E^{1/2}(\mu_2^2(\theta_k))E^{1/2}(\theta_k - \bar{\theta})^4 \\ &= n_T E^{1/2}(\mu_2^2(\theta_1))E^{1/2}(\theta_1 - \bar{\theta})^4 \\ &\leq CL, \end{aligned} \tag{4.62}$$

using assumptions (b)–(d). In addition, an application of Lemma 4.4.2 to the $\theta_k - \mu$ gives

$$V\left[\sum_{k=1}^{L} n_k(\theta_k - \bar{\theta})^2\right] = \left[\sum_{k=1}^{L} n_k^2(1 - n_T^{-1}n_k)^2\right][E(\theta_1 - \mu)^4 - \tau^4]$$

$$+ 2\left[\sum\sum_{1\le k\neq k'\le L} n_k^2 n_{k'}^2/n_T^2\right]\tau^4$$
$$= C[LE(\theta_1-\mu)^4+\tau^4], \tag{4.63}$$

using assumptions (a)–(d). Finally, using $E[\sum_{k=1}^{L} n_k(\bar{z}_k-\bar{z})^2|\boldsymbol{\theta}] = \sum_{k=1}^{L}(1-n_k n_T^{-1})\mu_2(\theta_k)$, one gets

$$V\left[\sum_{k=1}^{L} n_k(\bar{z}_k-\bar{z})^2\right] = E\left[V\left\{\sum_{k=1}^{L} n_k(\bar{z}_k-\bar{z})^2|\boldsymbol{\theta}\right\}\right] + V\left[\sum_{k=1}^{L}(1-n_k n_T^{-1})\mu_2(\theta_k)\right]. \tag{4.64}$$

An application of Lemma 4.4.2 with $\boldsymbol{W} = (\bar{z}_1,\ldots,\bar{z}_L)^T$ and $\boldsymbol{A}$ as defined in Lemma 4.4.2 now leads to

$$V\left[\sum_{k=1}^{L} n_k(\bar{z}_k-\bar{z})^2|\boldsymbol{\theta}\right]$$
$$= \sum_{k=1}^{L} n_k^2(1-n_T^{-1}n_k)^2[E(\bar{z}_k^4|\theta_k)-(E(\bar{z}_k^2|\theta_k))^2]$$
$$+ 2\sum\sum_{1\le k\neq k'\le L} n_k^2 n_{k'}^2 n_T^{-2} E(\bar{z}_k^2|\theta_k)E(\bar{z}_{k'}^2|\theta_{k'}). \tag{4.65}$$

But

$$E(\bar{z}_k^4|\theta_k) = n_k^{-3}[\mu_4(\theta_k)+6(n_k-1)\mu_2^2(\theta_k)],$$

and

$$E(\bar{z}_k^2|\theta_k) = V(\bar{z}_k|\theta_k) = n_k^{-1}\mu_2(\theta_k).$$

Hence, from (4.65) one gets

$$V\left[\sum_{k=1}^{L} n_k(\bar{z}_k-\bar{z})^2|\boldsymbol{\theta}\right] \le \sum_{k=1}^{L} n_k^{-1}[\mu_4(\theta_k)+6(n_k-1)\mu_2^2(\theta_k)] + 2\sum\sum_{1\le k\neq k'\le L} n_k^2 n_{k'}^2 \mu_2(\theta_k)\mu_2(\theta_{k'}). \tag{4.66}$$

Next, using $E[\mu_2^2(\theta_1)] \le E[\mu_4(\theta_1)]$, it follows from (4.64) and (4.66)

that under the assumptions (a) and (b),

$$V\left[\sum_{k=1}^{L} n_k(\bar{z}_k - \bar{z})^2\right] \leq CLE[\mu_4(\theta_1)]. \tag{4.67}$$

Now using assumptions (b) and (d), it follows from (4.61), (4.62), (4.63) and (4.64) that $V(SSB) \leq CL$, so that $V(MSB) \leq CL^{-1}$. This completes the proof of the lemma. □

We are now in a position to prove Theorem 4.4. Define

$$V_L = L^{-1}\sum_{k=1}^{L} f_k^2[(\hat{B}_k - B_k)(\bar{z}_k - m) - \hat{B}_k(\hat{m} - m)]^2. \tag{4.68}$$

Note that $r(\xi, e_{EB}) - r(\xi, e_B) = E(V_L)$. We prove the theorem by showing that $V_L \to 0$ in probability as $L \to \infty$, and then proving the uniform integrability in L of V_L.

To prove that $V_L \to 0$ in probability, use the inequality,

$$V_L \leq 2\left[max_{1\leq k\leq L}(\hat{B}_k - B_k)^2 L^{-1}\sum_{k=1}^{L}(\bar{z}_k - m)^2 + (\hat{m} - m)^2\right]. \tag{4.69}$$

Using Lemma 4.4.3 as well as Lemma 4.3.2, one gets

$$max_{1\leq k\leq L}|\hat{B}_k - B_k| \to 0 \text{ as } L \to \infty.$$

Moreover, using the arguments given in (4.34)–(4.37), one gets $\hat{m} \to m$ in probability. Thus, $V_L \to 0$ in probability.

To prove the uniform integrability of V_L, first use the inequality,

$$V_L \leq 2\left[L^{-1}\sum_{k=1}^{L}(\bar{z}_k - M)^2 + (\hat{m} - m)^2\right] \tag{4.70}$$

from (4.69). Note that

$$\begin{aligned} &E\left[L^{-1}\sum_{k=1}^{L}(\bar{z}_k - m)^2\right] \\ &\quad \leq E\left[L^{-1}\sum_{k=1}^{L}(\bar{z}_k - m)^4\right] \\ &\quad \leq 8L^{-1}\sum_{k=1}^{L} E[(\bar{z}_k - \theta_k)^4 + (\theta_k - m)^4]. \end{aligned} \tag{4.71}$$

But

$$L^{-1}\sum_{k=1}^{L} E(\bar{z}_k - \theta_k)^4 \leq \left[L^{-1}\sum_{k=1}^{L}(n_k^{-3} + 6n - [k]^{-2})\right] E\mu_4(\theta_1)$$
$$\leq (13/8)E\mu_4(\theta_1), \tag{4.72}$$

using assumption (a). Now, using assumptions (c) and (d), it follows from (4.71) and (4.72) that

$$sup_{L\geq 1} E\left[L^{-1}\sum_{k=1}^{L}(\bar{z}_k - m)^4\right] < \infty, \tag{4.73}$$

which proves the uniform integrability of the first term in the right-hand side of (4.69). Next, using (4.38), one gets the uniform integrability of $(\hat{m} - m)^2$. This completes the proof of the theorem.

In the case when $n_1 = \cdots = n_L = n$, one has

$$g = n(L-1);$$

$$\hat{\lambda}^{-1} = max[0, ((L-1)MSB/((L-1)MSW) - 1)];$$

$$\hat{B} = min[1, (L-3)MSW/((L-1)MSB)].$$

Then $\hat{\mu}^{EB}$ reduces to the analogue of the positive-part James–Stein estimator. The EB estimator of μ is then $\tilde{e} = (\tilde{e}_1, \ldots, \tilde{e}_L)^T$, where

$$\tilde{e}_k = \bar{z}_k - f_k\hat{B}(\bar{z}_k - \bar{z}), \tag{4.74}$$

which is the finite population analogue of Lindley's (Lindley, 1962) modification of the James–Stein estimator. The final main theorem of this section is as follows.

Theorem 4.5 *Suppose that $n_1 = \cdots, n_L = n$. Under the assumed model with the extra condition that the original y_i are iid, one has*

$$RSL(\xi, \tilde{e}, \tilde{\mu_0})$$
$$= L^{-1}[((L-3)/(L(n-1)))^2$$
$$\times E[(SSW)^2/SSB](B\sigma^2)^{-1} - (L-6)]. \tag{4.75}$$

If, in addition,

$$E[(SSW)^2/SSB] = E(SSW)^2/E(SSB) + O(1), \tag{4.76}$$

then $RSL(\xi; \tilde{e}, \tilde{\mu_0}) = O(L^{-1})$.

Remark 2. If normality is assumed, then $SSB \sim \sigma^2 B^{-1}\chi^2_{L-1}$. In this case, exact calculations give

$$E[(SSW)^2/SSB] = B\sigma^2 L(n-1)(L(n-1)+2)/(L-3). \tag{4.77}$$

Then from (4.75) and (4.77) it follows that

$$RSL(\xi;\tilde{e},\tilde{\mu}_0) = (3n-1-6L^{-1})/(L(n-1)). \tag{4.78}$$

It is also easy to see in this case that (4.76) holds. Similarly, (4.76) holds when (i)–(iii) is assumed, and $E(h^{-1}) < \infty$. The details appear in Lahiri (1986).

Proof. First note that since $n_1 = \cdots = n_L = n$, $B_1 = \cdots = B_L = B$. Hence,

$$RSL(\xi,\tilde{e},\tilde{\mu}) = \sum_{k=1}^{L} f_k^2 E[\hat{B}(\bar{z}_k - \bar{z}) - B(\bar{z}_k - m)]^2 \div \left[B\sigma^2 n^{-1}\sum_{k=1}^{L} f_k^2\right]. \tag{4.79}$$

Using the fact that $(\tilde{B}-B)(\bar{z}_k-\bar{z})$ has the same distribution for every $k = 1,\ldots,L$, and $(\tilde{B}-B)(\bar{z}_k-\bar{z})(\bar{z}-m)$ has the same distribution for every $k = 1,\ldots,L$, it follows that

$$\begin{aligned}
&E[\tilde{B}(\bar{z}_k-\bar{z}) - B(\bar{z}_k-m)]^2 \\
&\quad = E[(\tilde{B}-B)(\bar{z}_k-\bar{z}) - B(\bar{z}-m)]^2 \\
&\quad = E[(\tilde{B}-B)^2(\bar{z}_k-\bar{z})^2 + B^2(\bar{z}-m)^2 \\
&\quad - 2(\tilde{B}-B)B(\bar{z}-m)(\bar{z}_k-\bar{z})] \\
&\quad = E\Bigg[(\tilde{B}-B)^2 L^{-1}\sum_{k=1}^{L}(\bar{z}_k-\bar{z})^2 + B\sigma^2 L^{-1}n^{-1} \\
&\quad - 2(\tilde{B}-B)B(\bar{z}-m)L^{-1}\sum_{k=1}^{L}(\bar{z}_k-\bar{z})\Bigg] \\
&\quad = (Ln)^{-1}[E((\tilde{B}-B)^2 SSB) + B\sigma^2].
\end{aligned} \tag{4.80}$$

Now writing $d = (L-3)/(L(n-1))$, and using Lemma 4.4.1, one gets

$$\begin{aligned}
&E(\tilde{B}-B)^2 SSB \\
&\quad = d^2 E[(SSW)^2/SSB] - 2dBE(SSW) + B^2 E(SSB)
\end{aligned}$$

$$= d^2E[(SSW)^2/SSB] - 2(L-3)B\sigma^2 + (L-1)B\sigma^2$$
$$= d^2E[(SSW)^2/SSB] - (L-5)B\sigma^2. \tag{4.81}$$

Equation (4.75) follows now from (4.79)–(4.81). Next note that for $n_1 = \cdots = n_L = n$,

$$\begin{aligned} V(SSW) &= L[(n-1)^2n^{-1}E\mu_4(\theta_1) \\ &+ (n-1)(n^2-2n+3)n^{-1}E\mu_2^2(\theta_1) \\ &- (n-1)^2\sigma^4]. \end{aligned} \tag{4.82}$$

In addition, from Lemma 4.4.1, $E(SSW) = L(n-1)\sigma^2$. Hence, if (4.76) holds, then from (4.81) and (4.82), one gets

$$E[(SSW)^2/SSB] = L(n-1)^2B\sigma^2 + O(1). \tag{4.83}$$

Substitution of (4.83) in (4.75) now yields

$$\begin{aligned} RSL(\xi;\tilde{e},\tilde{\mu}_0) &= L^{-1}[(L-3)^2L^{-1} - (L-6)] + O(L^{-1}) \\ &= O(L^{-1}). \end{aligned} \tag{4.84}$$

Further,

$$\begin{aligned} E(SSW)^2 &= V(SSW) + (ESSW)^2 \\ &= L(n-1)[(n-1)n^{-1}E\mu_4(\theta_1) \\ &+ (n^2-2n+3)n^{-1}E(\mu_2^2(\theta_1)) \\ &+ (L-1)(n-1)\sigma^4]. \end{aligned} \tag{4.85}$$

If $n_1 = \cdots = n_L = n$, then from Lemma 4.4.1,

$$E(SSB) = (L-1)(\sigma^2 + n\tau^2) = (L-1)\sigma^2B^{-1}. \tag{4.86}$$

This completes the proof. □

Next in this section, we compare numerically the Bayes risks of the EB estimator $\hat{\mu}^{EB}$ and the classical estimator $\tilde{\mu}_0$ of μ under squared error loss. The underlying distribution is assumed to be normal. Consider the case in which $L = 5$ and $n_1 = 1$, $n_2 = 2$, $n_3 = 3$, $n_4 = 5$ and $n_5 = 10$. As before, let μ_k denote the true mean of the kth stratum. Two cases are reported here: (a) $\mu_1 = \ldots = \mu_5 = 0$ and $m = 5$, and (b) $\mu_1 = 0$, $\mu_2 = 2$, $\mu_3 = 4$, $\mu_4 = 6, \mu_5 = 3$ and $m = 0$. In all the cases, we take $f_k = 0.9$ for all $k = 1, \ldots, 5$ and $\sigma^2 = 1$. The Bayes risks are calculated as functions of τ^2. The random normal deviates are generated using the RANNOR-generating function of the SAS library. The simulated risks are the average losses after 1000 repetitions of the experiment. The Monte Carlo findings are reported in Tables 4.1 and 4.2

Table 4.1 *The Bayes risks of* $\tilde{\mu}_0$, $\hat{\mu}^{EB}$, $\hat{\mu}^{B}$ *and the percentage risk improvement (PCTIMP) of* $\hat{\mu}^{EB}$ *over* $\tilde{\mu}_0$ *in Case (a).*

τ	$r(\xi, \tilde{\mu}_0)$	$r(\xi, \hat{\mu}^{EB})$	PCTIMP	$r(\xi, \hat{\mu}^{B})$
0.1	0.3840	0.1656	56.8750	18.7472
0.3	0.3840	0.1971	48.6719	11.9790
0.5	0.3840	0.2381	37.9948	6.9956
0.8	0.3840	0.2885	24.8698	3.2847
1.0	0.3840	0.3125	18.6198	2.1172
1.5	0.3840	0.3474	9.5313	0.9282
2.0	0.3840	0.3635	5.3385	0.5841
2.5	0.3840	0.3717	3.2031	0.4662

Table 4.2 *The Bayes risks of* $\tilde{\mu}_0$, $\hat{\mu}^{EB}$, $\hat{\mu}^{B}$ *and the percentage risk improvement (PCTIMP) of* $\hat{\mu}^{EB}$ *over* $\tilde{\mu}_0$ *in Case (b).*

τ	$r(\xi, \tilde{\mu}_0)$	$r(\xi, \hat{\mu}^{EB})$	PCTIMP	$r(\xi, \hat{\mu}^{B})$
0.3	0.3840	0.3757	2.1614	5.3434
0.5	0.3840	0.3760	2.0833	2.5559
0.8	0.3840	0.3794	1.1979	1.0072
1.0	0.3840	0.3816	0.6250	0.6637
1.5	0.3840	0.3837	0.0781	0.4152
2.0	0.3840	0.3831	0.2344	0.3739
2.5	0.3840	0.3827	0.3385	0.3678

The subjective Bayes estimator makes a wrong guess at the prior means for all the five strata in Case (a), and four out of the five strata in Case (b). It is not surprising, therefore, that it performs very poorly in both these cases, and this is clearly reflected in the two tables. The EB estimator compares most favourably with the sample mean in Case (a) when exchangeability holds among the different strata. The PCTIMP decreases naturally with τ since higher values of τ indicate greater uncertainty about the prior. The interesting feature is the performance of the EB estimator in Case (b) when exchangeability does not hold among the strata. Even in this case, $\hat{\mu}^{EB}$ performs slightly better than $\tilde{\mu}_0$. This indicates

clearly the robustness of the EB estimator. Other examples worked out in Lahiri (1986) indicate the same feature.

Next the binomial example is considered. In this case, the y_i that belong to stratum k are iid Bin(1, θ_k). Also, the θ_k are assumed to be iid Beta(p, q). Various choices of p and q are considered. For comparison with other rival estimators, some of which are proposed only for the case $n_1 = \cdots = n_L$, we consider this balanced situation. In this case

$$\hat{B} = min\left\{1, (L-3)\sum_{k=1}^{L}\bar{z}_k(1-\bar{z}_k)/\left[L(n-1)\sum_{k=1}^{L}(\bar{z}_k-\bar{z})^2\right]\right\}.$$

Two rival estimators can be developed along the lines of Morris (1983c) and Robbins (1983) . Following Morris define an estimator B_M of B by

$$\begin{aligned} B_M &= \left[(L-3)\bar{z}(1-\bar{z})/\left\{(n-1)\sum_{k=1}^{L}(\bar{z}_k-\bar{z})^2\right\}\right] \\ &- (L-1)/(L(n-1)). \end{aligned}$$

Since B_M can take negative values or can exceed 1, we modify it by $\hat{B}_M = 0$, B_M or 1 according as $B_M < 0$, $0 \leq B_M \leq 1$ or $B_M > 1$, respectively. The corresponding EB estimator of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}^M = (\hat{\mu}_1^M, \ldots, \hat{\mu}_L^M)^T$, where

$$\hat{\mu}_k^M = \bar{z}_k - f_k\hat{B}_M(\bar{z}_k - \bar{z}),\ k = 1, \ldots, L. \tag{4.87}$$

Again, following Example 3 of Robbins (1983), define an estimator B_R of B by

$$B_R = \left[(L-1)\bar{z}(1-\bar{z})/\left\{(n-1)\sum_{k=1}^{L}(\bar{z}_k-\bar{z})^2\right\}\right] - (n-1)^{-1}.$$

Since B_R can also take negative values or can exceed 1, it is modified by $\hat{B}_R = 0$, B_R or 1, according as $B_R < 0$, $0 \leq B_R \leq 1$ or $B_R > 1$, respectively. The corresponding EB estimator of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}^R$, where

$$\hat{\mu}_k^R = \bar{z}_k - f_k\hat{B}_R(\bar{z}_k - \bar{z}),\ k = 1, \ldots, L. \tag{4.88}$$

Monte Carlo simulations are performed to compare the Bayes risks of $\hat{\boldsymbol{\mu}}^{EB}$, $\hat{\boldsymbol{\mu}}^M$ and $\hat{\boldsymbol{\mu}}^R$ for various choices of (p, q). In all these cases $L = 5, n = 4$, and $f_k = 0.9$ for all k. The simulated Bayes risks are the losses averaged over 1000 repetitions of the experiment for each (p, q). First, the θ_k's are generated from a beta(p, q)

Table 4.3 *The Bayes risks of* $\tilde{\mu}_0$, $\hat{\mu}^{EB}$, $\hat{\mu}^M$ *and* $\hat{\mu}^R$, *and the percentage risk improvements.*

p	q	$r(\xi, \tilde{\mu}_0)$	$r(\xi, \hat{\mu}^{EB})$	$r(\xi, \hat{\mu}^M)$	$r(\xi, \hat{\mu}^R)$
1	1	0.0374	0.0329	0.0359	0.0344
			(12.03)	(4.01)	(8.02)
0.5	0.5	0.0281	0.0262	0.0278	0.0278
			(6.76)	(1.07)	(1.07)
0.5	0.7	0.0298	0.0275	0.0294	0.0293
			(7.72)	(1.34)	(1.68)
0.5	0.4	0.0181	0.0140	0.0156	0.0142
			(22.22)	(13.81)	(21.54)
2	0.5	0.0254	0.0218	0.0236	0.0230
			(14.17)	(7.09)	(9.45)
2	1	0.0375	0.0312	0.0346	0.0319
			(16.80)	(7.73)	(14.93)
2	2	0.0445	0.0355	0.0399	0.0353
			(20.22)	(10.34)	(20.67)
2	4	0.0426	0.0319	0.0363	0.0305
			(25.12)	(14.79)	(28.40)
3	1	0.0334	0.0267	0.0297	0.0269
			(20.06)	(11.08)	(19.46)
3	3	0.0475	0.0356	0.0405	0.0340
			(25.05)	(14.74)	(28.42)
3	4	0.0476	0.0347	0.0398	0.0326
			(27.10)	(16.39)	(31.51)

distribution. Then the $y_i, i \in str_k$ are generated from a binomial(1, θ_k) distribution, $k = 1, \ldots, L$. The simulated Bayes risks of $\tilde{\mu}_0$, $\hat{\mu}^{EB}$, $\hat{\mu}^M$ and $\hat{\mu}^R$ are reported in Table 4.3. The percentage risk improvemements of $\hat{\mu}^{EB}$, $\hat{\mu}^M$ and $\hat{\mu}_R$ over $\tilde{\mu}_0$ are given within parentheses.

Various types of beta distributions are considered in Table 4.3: uniform ($p = q = 1$), symmetric U-shaped ($p = q < 1$), symmetric bell-shaped ($p = q > 1$), J-shaped ($p > 1$, $q < 1$), inverse J-shaped ($p < 1$, $q > 1$), and nonsymmetric bell-shaped ($p > 1$, $q > 1$). In all cases $\hat{\mu}^{EB}$ does better than $\tilde{\mu}_0$ and $\hat{\mu}^M$, and there is no clear choice between $\hat{\mu}^{EB}$ and $\hat{\mu}^R$ even though $\hat{\mu}^R$ does use the addi-

tional information regarding the variance structure of binomial. In any case $\hat{\boldsymbol{\mu}}^{EB}$ seems to be fairly robust in this case. The small differences in the Bayes risks can be accounted for by the appearance of the normalizing L^{-1} in the loss. This is why the percentage risk improvement figures are more meaningful.

Next consider the Poisson case. Again $n_1 = \ldots = n_L = n$. Following Morris (1983), define

$$\hat{B}^{MO} = min\left[1, (L-3)\bar{z}/\left\{n\sum_{k=1}^{L}(\bar{z}_k - \bar{z})^2\right\}\right]$$

as an estimator of B. A similar estimator

$$\hat{B}^{L} = min\left[1, (L-1)\bar{z}/\left\{n\sum_{k=1}^{L}(\bar{z}_k - \bar{z})^2\right\}\right]$$

was proposed by Leonard (1976). The corresponding EB estimators are $\hat{\boldsymbol{\mu}}^{MO} = (\hat{\mu}_1^{MO}, \cdots, \hat{\mu}_L^{MO})^T$ and $\hat{\boldsymbol{\mu}}^{L} = (\hat{\mu}_1^{L} \ldots, \hat{\mu}_L^{L})^T$ where

$$\hat{\mu}_k^{MO} = \bar{z}_k - f_k\hat{B}^{MO}(\bar{z}_k - \bar{z}),\ k = 1, \ldots, L; \tag{4.89}$$

and

$$\hat{\mu}_k^{L} = \bar{z}_k - f_k\hat{B}^{L}(\bar{z}_k - \bar{z}),\ k = 1, \ldots, L. \tag{4.90}$$

For Monte Carlo simulations we first generate the θ_k from a gamma(α, p), that is a distribution with pdf

$$f_{\alpha,p}(x) = [exp(-\alpha x)\alpha^p x^{p-1}/\Gamma(p)]I_{[x>0]}.$$

Then we generate y_i such that for $i \in str_k$, y_i are iid Poisson(θ_k). Once again, we take $L = 5$, $n = 4$ and $f_k = .9$ for all k. The simulated Bayes risks are found similarly to the binomial cases. Various choices of p and α are considered. Table 4.4 reports the Bayes risks of $\tilde{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\mu}}^{EB}$, $\hat{\boldsymbol{\mu}}^{MO}$ and $\hat{\boldsymbol{\mu}}^{L}$. Also, the percentage risk improvement of the last three over $\tilde{\boldsymbol{\mu}}_0$ are given within parentheses. The cases $p = 1$, $p = .5$, and $p = 3$ correspond respectively to the exponential, negatively skewed gamma, and the positively skewed gamma distributions.

It follows from Table 4.4 that in most cases $\hat{\boldsymbol{\mu}}^{MO}$ is the winner, but $\hat{\boldsymbol{\mu}}^{EB}$ finishes a close second (within 1.5% of the percentage risk improvement). It seems also that $\hat{\boldsymbol{\mu}}^{L}$ is the appropriate estimator when the mean p/α of the gamma distribution is small. Overall, $\hat{\boldsymbol{\mu}}^{EB}$ performs quite respectably in comparison with other estimators that use more information about the structure of the superpopulation.

Table 4.4 *The Bayes risks of* $\tilde{\mu}_0$, $\hat{\mu}^{EB}$, $\hat{\mu}^{MO}$ *and* $\hat{\mu}^{L}$ *and the percentage risk improvements.*

p	α	$r(\xi, \tilde{\mu}_0)$	$r(\xi, \hat{\mu}^{EB})$	$r(\xi, \hat{\mu}^{MO})$	$r(\xi, \hat{\mu}^{L})$
.5	0.1	1.1263	1.1204	1.1178	1.1307
			(0.52)	(0.75)	(−0.39)
	0.5	0.2258	0.2186	0.2169	0.2242
			(3.19)	(3.94)	(0.71)
	1	0.1134	0.1057	0.1042	0.1076
			(6.79)	(8.11)	(5.11)
	4	0.0279	0.0220	0.0216	0.0210
			(21.15)	(22.58)	(24.73)
1	0.1	2.2509	2.2347	2.2316	2.2527
			(0.72)	(0.86)	(−0.08)
	0.5	0.4514	0.4356	0.4323	0.4438
			(3.50)	(4.23)	(1.68)
	1	0.2243	0.2088	0.2065	0.2126
			(6.91)	(7.93)	(5.22)
	4	0.0562	0.0448	0.0440	0.0420
			(20.28)	(21.71)	(25.27)
3	0.1	6.7506	6.6945	6.6882	6.7340
			(0.83)	(0.92)	(0.24)
	0.5	1.3512	1.3007	1.2929	1.3179
			(3.74)	(4.31)	(2.46)
	1	0.6761	0.6254	0.6217	0.6345
			(7.50)	(8.05)	(6.15)
	4	0.1666	0.1321	0.1300	0.1230
			(20.70)	(21.97)	(26.17)

Other estimators vary according to the choice of a specific distribution, but $\hat{\mu}^{EB}$ does not. Moreover, for Dirichlet process priors, posterior linearity holds so that $\hat{\mu}^{EB}$ is still appropriate, but other EB estimators are not. We recommend that unless one is very confident about the use of a specific prior, it is always safe to use $\hat{\mu}^{EB}$ as the estimator of the finite population mean μ rather than any other EB estimator.

4.5 Multistage sampling

The previous section considered simultaneous estimation of means from several strata when no distributional assumptions were involved. The results are generalized in this section to two-stage sampling. Suppose that each unit within a stratum (called a primary unit) can be divided into a smaller number of units or subunits. A sample of primary units is selected at the first stage. If subunits within a primary unit give similar results, it seems uneconomical to measure them all. A common practice is then to select and measure a sample of subunits from the chosen primary unit. The technique is called **subsampling** or **two-stage sampling**. A very common example is household surveys where counties are used as different strata. A sample of blocks within each of the counties is drawn at the first stage, and a sample of dwellings from within the chosen blocks is drawn at the second stage.

Scott and Smith (1969) carried out a Bayesian analysis for two-stage sampling within a single stratum assuming a normal prior. Their results were generalized to three-stage sampling by Malec and Sedransk (1985). However, in keeping with the previous section, we dispense with the normality assumption, and assume posterior linearity instead. We shall first derive Bayes estimators of the means of the strata assuming posterior linearity.

Suppose there are L strata, where the kth stratum contains N_{kj} secondary units or subunits. As before, we denote by y_i the characteristic associated with the ith unit. The following hierarchical model is assumed.

(A) Consider a unit i belonging to the jth primary unit within the kth stratum. Its distribution depends only on θ_{kj}, with $E(y_i|\theta_{kj}) = \theta_{kj}$ and $V(y_i|\theta_{kj}) = \mu_2(\theta_{kj})$, $j = 1, \ldots, M_k$, $k = 1, \ldots, L$.

(B) Conditional on q_k, θ_{kj} are iid with a distribution depending only on q_k with $E(\theta_{kj}|q_k) = q(k)$ and $V(\theta_{kj}|q_k) = \mu_2(q_k)$.

(C) The q_k are iid with $E(q_k) = \nu$ and $V(q_k) = \delta^2$.

Assume also that

$$\sigma^2 = E(\mu_2(\theta_{kj})),\ \tau^2 = E(\mu_2(q_k)).$$

As an example where assumptions (A)–(C) hold, consider the one-way variance components model where for unit i belonging to

the jth primary unit within the kth stratum,

$$y_i = \nu + \xi_k + \eta_{kj} + e_i, \tag{4.91}$$

where the ξ_k, η_{kj} and the e_i are mutually independent. Also, the e_i are iid with $E(e_i) = 0$, $V(e_i) = \sigma^2$; η_{kj} are iid with $E(\eta_{kj}) = 0$, $V(\eta_{kj}) = \tau^2$; ξ_k are iid with $E(\xi_k) = 0$, $V(\xi_k) = \delta^2$.

As before, our target is to estimate the vector of the population means $\mu_k = N_k^{-1} \sum_{i \in str_k} y_i$, $k = 1, \ldots, L$, under the loss

$$L(a, \mu) = L^{-1} \sum_{k=1}^{L} (a_k - \mu_k)^2. \tag{4.92}$$

For the kth stratum, a sample of m_k primary units are taken. For the jth primary unit within the kth stratum, a sample of n_{kj} secondary units are taken. Without loss of generality, we denote the selected primary units by $1, \ldots, m_k$. Similarly, the selected secondary units within the jth primary unit (if selected) are denoted by $1, \ldots, n_{kj}$. We shall use the notations $n_k = \sum_{j=1}^{m_k}$, $N_k = \sum_{j=1}^{m_k} N_{kj}$, $m_T = \sum_{k=1}^{L} M_k$.

As in the previous section, we assume posterior linearity, that is

$$E(\theta_{kj}|z) = \sum_{l=1}^{m_k} \sum_{\{i \in s, i \in psu\ l \text{ in } str_k\}} a_{kli}^j z_i + a_{k0}^j. \tag{4.93}$$

The coefficients a_{kli}^j and a_{k0}^j are obtained by minimizing

$$E\left(\theta_{kj} - \sum_{l=1}^{m_k} \sum_{\{i \in s, i \in psu\ l \text{ in } str_k\}} a_{kli}^j - a_{k0}^j\right)^2 \tag{4.94}$$

with respect to a_{kli}^j and a_{k0}^j.

To achieve the desired minimization, one uses the assumptions (A)–(C) to get

$$E(y_i) = \nu; \tag{4.95}$$

$$V(y_i) = \sigma^2 + \tau^2 + \delta^2; \tag{4.96}$$

$$Cov(y_i, y_{i'}) = \tau^2 + \delta^2,\ i, i' \text{ in the same psu}; \tag{4.97}$$

$$Cov(y_i, y_{i'}) = \delta^2,\ i, i' \text{ in different psu}; \tag{4.98}$$

$$Cov(\theta_{kj}, y_i) = \tau^2 + \delta^2, i \in \text{psu}j; \tag{4.99}$$

$$Cov(\theta_{kj}, y_i) = \delta^2, i \notin \text{psu}j. \tag{4.100}$$

Hence, writing z_k as the vector of sampled observations from the

kth stratum, $d^2 = \delta^2 + \nu^2$ and $h^2 = \tau^2 + d^2$, Theorem 2.1 of Goldstein (1975) leads to

$$E(\theta_{kj}) = (1,\ \boldsymbol{z}_k^T)\boldsymbol{D}_{k0}^{-1}\boldsymbol{b}_{k0}, \tag{4.101}$$

where $\boldsymbol{D}_{k0}$ is given by

$$\begin{bmatrix} 1 & \nu\mathbf{1}_{n_1}^T & \cdots & \nu\mathbf{1}_{n_L}^T \\ \nu\mathbf{1}_{n_1} & \sigma^2\boldsymbol{I}_{n_1} + h^2\boldsymbol{J}_{n_1} & \cdots & d^2\mathbf{1}_{n_1}\mathbf{1}_{n_L}^T \\ \nu\mathbf{1}_{n_L} & d^2\mathbf{1}_{n_L}\mathbf{1}_{n_1}^T & \cdots & \sigma^2\boldsymbol{I}_{n_L} + h^2\boldsymbol{J}_{n_L} \end{bmatrix}; \tag{4.102}$$

and

$$\boldsymbol{b}_{k0}^T = [\nu(\delta^2+\nu^2)\mathbf{1}_{n_1}^T \cdots (\tau^2+\delta^2+\nu^2)\mathbf{1}_{n_k}^T \cdots (\delta^2+\nu^2)\mathbf{1}_{n_L}^T]. \tag{4.103}$$

From (4.102) and (4.103), it is easy to observe that the right- hand side of (4.101) depends on $\boldsymbol{z}_k$ only through $\bar{\boldsymbol{z}}_k = (\bar{z}_{k1}, \ldots, \bar{z}_{km_k})^T$. Now,

$$E(\theta_{kj}|\bar{\boldsymbol{z}}) = \sum_{l=1}^{m} e_{kl*}^j \bar{z}_{kl} + e_{k0*}^j, \tag{4.104}$$

where the coefficients e_{kl*}^j and e_{k0*}^j are obtained by minimizing

$$\sum_{j=1}^{m} E\left(\theta_{kj} - \sum_{l=1}^{m} e_{kl}^j \bar{z}_{kl} - e_{k0}^j\right)^2$$

with respect to e_{kl}^j and e_{k0}^j. We shall now obtain in the following theorem the Bayes estimators of the θ_{kj} under squared error loss. First, we need a few notations.

Let $\lambda_1 = \sigma^2/\tau^2$, $\lambda_2 = \sigma^2/\delta^2$. Also, let $B_{kj} = \lambda_1/(\lambda_1 + n_{kj})$, and $u_k = (\lambda_2 + \lambda_1 \sum_{j=1}^{m_k}(1 - B_{kj}))^{-1}(\lambda_2\nu + \lambda_1 \sum_{j=1}^{m_k}(1 - B_{kj})\bar{z}_{kj})$. The following theorem is now proved.

Theorem 4.6 *Under the model given in (A)–(C), for* $1 \le j \le m_k$, $E(\theta_{kj}|\bar{\boldsymbol{z}}) = (1 - B_{kj})\bar{z}_{kj} + B_{kj}u_k$, *while for* $m_k + 1 \le j \le M_k$, $E(\theta_{kj}|\bar{\boldsymbol{z}}) = u_k$.

Proof. First a few preliminary calculations are needed. We have

$$E(\bar{z}_{kj}) = \nu; \tag{4.105}$$

$$V(\bar{z}_{kj}) = n_{kj}^{-1}\sigma^2 + \tau^2 + \delta^2; \tag{4.106}$$

$$Cov(\bar{z}_{kj}, \bar{z}_{kj'}) = \delta^2 \ (j \ne j'); \tag{4.107}$$

$$Cov(\theta_{kj}, \bar{z}_{kj}) = \tau^2 + \delta^2; \tag{4.108}$$

$$Cov(\theta_{kj}, \bar{z}_{kj'}) = \delta^2 \ (j \ne j'). \tag{4.109}$$

Let $\boldsymbol{A}_k = Diag(n_{k1}^{-1}\sigma^2 + \tau^2, \ldots, n_{km_k}^{-1}\sigma^2 + \tau^2)$, and $\boldsymbol{e}_{m_k}^{(j)}$ the m_k-component column vector with 1 in the jth position, and zeroes elsewhere. Then using Theorem 2.1 of Goldstein (1975) once again, it follows that for $1 \leq j \leq m_k$,

$$E(\theta_{kj}|\bar{\boldsymbol{z}}) = (1\ \bar{\boldsymbol{z}}_k^T)\boldsymbol{D}_k^{-1}\boldsymbol{b}_k^{(j)}, \tag{4.110}$$

where

$$\boldsymbol{D}_k = \begin{bmatrix} 1 & \nu \mathbf{1}_{m_k}^T \\ \nu \mathbf{1}_{m_k} & \boldsymbol{A}_k + d^2 \boldsymbol{J}_{m_k} \end{bmatrix}, \tag{4.111}$$

and

$$\boldsymbol{b}_k^{(j)} = \begin{bmatrix} \nu \\ d^2 \mathbf{1}_{m_k} + \tau^2 \boldsymbol{e}_{m_k}^{(j)}. \end{bmatrix}. \tag{4.112}$$

It follows after some heavy algebra that $\boldsymbol{D}_k^{-1}\boldsymbol{b}_k^{(j)}$ is given by

$$\begin{bmatrix} \nu B_{kj}\lambda_2(\lambda_2 + \lambda_1(1 - B_{kj}))^{-1} \\ \lambda_1 B_{kj}(\lambda_2 + \lambda_1(1 - B_{kj}))^{-1}(\mathbf{1}_{m_k} - \boldsymbol{B}_k) + (1 - B_{kj})\boldsymbol{e}_{m_k}^{(j)} \end{bmatrix} \tag{4.113}$$

where $\boldsymbol{B}_k = (B_{k1}, \ldots, B_{km_k})^T$. From (4.110) and (4.112), some algebraic manipulations lead to

$$E(\theta_{kj}|\bar{\boldsymbol{z}}) = (1 - B_{kj})\bar{z}_{kj} + B_{kj}u_k \tag{4.114}$$

for $1 \leq j \leq m_k$. For $m_k + 1 \leq j \leq M_k$,

$$E(\theta_{kj}|\bar{\boldsymbol{z}}) = (1\ \bar{\boldsymbol{z}}_k^T)\boldsymbol{D}_k^{-1}(\nu\ d^2 \mathbf{1}_{m_k}^T)^T = u_k. \tag{4.115}$$

This completes the proof of the theorem. □

The above theorem is now used to derive the Bayes estimators of the stratum means μ_k.

Theorem 4.7 *Under the model given in (A)–(C), assuming squared error loss,*

$$E(\mu_k|\boldsymbol{z}) = u_k + \sum_{j=1}^{m_k} r_{kj}(1 - f_{kj}B_{kj})(\bar{z}_{kj} - u_k), \tag{4.116}$$

where $r_{kj} = N_{kj}/N_k$ *and* $f_{kj} = (N_{kj} - n_{kj})/N_{kj}$.

Proof. Recall that $\mu_k = N_k^{-1}\sum_{i \in str_k} y_i$. Hence, from Theorem 4.6, one gets

$$E(\mu_k|\boldsymbol{z}) \quad = \quad N_k^{-1}\Bigg[\sum_{j=1}^{m_k} \sum_{\{i \in s,\ i \in psu\ j \text{ in } str_k\}} z_i$$

$$
\begin{aligned}
&+ \sum_{j=1}^{m_k} \sum_{\{i\in\bar{s},\, i\in \textit{psu } j \textit{ in } \textit{str}_k\}} E(y_i|z) \\
&+ \sum_{j=m_k+1}^{M_k} \sum_{\{i\in\bar{s}, i\in \textit{psu} j \textit{ in } \textit{str}_k\}} E(y_i|z)\Bigg] \\
&= N_k^{-1}\Bigg[\sum_{j=1}^{m_k} n_{kj}\bar{z}_{kj} + \sum_{j=1}^{m_k}(N_{kj}-n_{kj})E(\theta_{kj}|z) \\
&+ \sum_{j=m_k+1}^{M_k} N_{kj}E(\theta_{kj}|z)\Bigg] \\
&= \sum_{j=1}^{m_k}(1-f_{kj}r_{kj})\bar{z}_{kj} \\
&+ \sum_{j=1}^{m_k} f_{kj}r_{kj}((1-B_{kj})\bar{z}_{kj} + B_{kj}u_k) \\
&+ \left(1-\sum_{j=1}^{m_k} r_{kj}\right)u_k \\
&= u_k + \sum_{j=1}^{m_k} r_{kj}(1-f_{kj}B_{kj})(\bar{z}_{kj}-u_k). \qquad (4.117)
\end{aligned}
$$

This completes the proof of the theorem. □

Remark 1. There is an interesting interpretation of u_k. We rewrite u_k as

$$
\begin{aligned}
u_k &= \left(\delta^{-2}\nu + \tau^{-2}\sum_{j=1}^{m_k}(1-B_{kj})\bar{z}_{kj}\right) \\
&\div \left(\delta^{-2} + \tau^{-2}\sum_{j=1}^{m_k}(1-B_{kj})\right). \qquad (4.118)
\end{aligned}
$$

Next we observe that $E(q_k) = \nu$ and $V(q_k) = \delta^2$. Also,

$$E[V(\bar{z}_{kj}|q_k)] = n_{kj}^{-1}\sigma^2 + \tau^2 = \tau^2(1-B_{kj})^{-1}.$$

Hence, from (4.118), u_k is a weighted average of the prior mean ν and sampled primary unit means $\bar{z}_{k1}, \ldots, \bar{z}_{km_k}$. The weights are reciprocals of $V(q_k)$, $E[V(\bar{z}_{k1}|q_k)]$, ..., and $E[V(\bar{z}_{km_k}|q_k)]$. It is

also clear from (4.117) that the Bayes estimator of μ_k shrinks the sample primary unit means $\bar{z}_{kj}$ to the pooled mean u_k.

Remark 2. When $M_k = 1$ for all $k = 1, \ldots, l$, that is each stratum has a single primary unit, then $\delta^2 = 0$, and so $u_k = \nu$ for all $k = 1, \ldots, L$. Then

$$E(\mu_k|\boldsymbol{z}) = N_{k1}^{-1}[n_{k1}\bar{z}_{k1} + (N_{k1} - n_{k1})((1 - B_{k1})\bar{z}_{k1} + B_{k1})\nu], \tag{4.119}$$

which is the same as the estimator obtained in the previous section with some changes in notation.

Remark 3. Scott and Smith (1969) considered two-stage estimation of the finite population mean under normal prior in the case of a single stratum. In order to make the results of this section comparable to theirs, consider the case when $\delta^2 = 0$. Then $u_1 = \nu$, and the expression given in (4.117) simplifies to

$$E(\mu_1|\boldsymbol{z}) = \nu + \sum_{j=1}^{m_1} r_{1j}(1 - f_{1j}B_{1j})(\bar{z}_{1j} - \nu). \tag{4.120}$$

With our notations, writing

$$\tilde{\nu} = \sum_{j=1}^{m_1}(1 - B_{1j})\bar{z}_{1j} / \sum_{j=1}^{m_1}(1 - B_{1j}),$$

one gets

$$e_B^* = \tilde{\nu} + \sum_{j=1}^{m_1} r_{1j}(1 - f_{1j}B_{1j})(\bar{z}_{1j} - \tilde{\nu}). \tag{4.121}$$

The only difference between the two estimators given in (4.120) and (4.121) is that the former involves ν, while the latter involves $\tilde{\nu}$. We shall see later in this section how (4.121) is obtainable from (4.120) from empirical Bayes considerations.

First, a general empirical Bayes methodology is developed in the context of two-stage sampling within strata. Initially, it is assumed that all the variance components σ^2, τ^2 and δ^2 are known, but ν is unknown, and needs to be estimated from the data. Later, we dispense with the assumption of known variance components.

Recall the definition of $\boldsymbol{D}_k$ as given in (4.111). Since $E(\bar{z}_k) = \nu \mathbf{1}_{m_k}$ and $V(\bar{z}_k) = \boldsymbol{D}_k$, in the absence of any distributional as-

sumption on the prior, the BLUE of ν is obtained by minimizing

$$Q = \sum_{k=1}^{L} (\bar{z}_k - \nu \mathbf{1}_{m_k})^T \boldsymbol{D}_k^{-1} (\bar{z}_k - \nu \mathbf{1}_{m_k}) \tag{4.122}$$

with respect to ν. Write $\boldsymbol{B}_k = (B_{k1}, \ldots, B_{km_k})^T$. Then, using the standard matrix inversion formula,

$$\boldsymbol{D}_k^{-1} = \tau^2 \Bigg[Diag(1 - B_{k1}, \ldots, 1 - B_{km_k}) - \lambda_1 \Bigg(\lambda_2 + \lambda_1 \sum_{j=1}^{m_k} (1 - B_{kj}) \Bigg)^{-1} (\mathbf{1}_{m_k} - \boldsymbol{B}_k)(\mathbf{1}_{m_k} - \boldsymbol{B}_k)^T \Bigg].$$

Next writing $g_k = \lambda_2 (\lambda_2 + \lambda_1 \sum_{j=1}^{m_k} (1 - B_{kj}))^{-1}$, and noting that $\lambda_2^{-1} \lambda_1 \sum_{j=1}^{m_k} (1 - B_{kj}) g_k = 1 - g_k$, it follows that the BLUE of ν is given by

$$\nu^* = \left[\sum_{k=1}^{L} g_k \sum_{j=1}^{m_k} (1 - B_{kj}) \right]^{-1} \left[\sum_{k=1}^{L} g_k \sum_{j=1}^{m_k} (1 - B_{kj}) \bar{z}_{kj} \right]. \tag{4.123}$$

In addition, if one assumes normality, ν^* is also the MLE of ν.

An EB estimator of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^T$ is then given by

$$e_{EB}^* = (e_{EB}^1, \ldots, e_{EB}^L)^T,$$

where

$$e_{EB}^k = u_k^* + \sum_{j=1}^{m_k} r_{kj} ((1 - f_{kj} B_{kj})(\bar{z}_{kj} - u_k^*)), \tag{4.124}$$

u_k^* being obtained by the replacement of ν by ν^* in (4.118).

Remark 4. The estimator ν^* has an interesting interpretation. Writing

$$\bar{z}_{wk} = \sum_{j=1}^{m_k} (1 - B_{kj}) \bar{z}_{kj} / \sum_{j=1}^{m_k} (1 - B_{kj})$$

and $c_k = g_k \sum_{j=1}^{m_k} (1 - B_{kj})$, it follows from (4.123) that ν^* can be alternatively expressed as $\nu^* = \sum_{k=1}^{L} c_k \bar{z}_{wk} / \sum_{k=1}^{L} c_k$. Note that $\bar{z}_{wk}$ is itself a weighted average of the primary unit sample means $\bar{z}_{k1}, \ldots, \bar{z}_{kL}$. This follows from Remark 1 of this section. Next we show that $V(\bar{z}_{wk})$ is inversely proportional to c_k.

Theorem 4.8 $V(\bar{z}_{wk}) = \tau^2 c_k^{-1}$.

Proof.

$$V(\bar{z}_{wk})=\left[\sum_{j=1}^{m_k}(1-B_{kj})\right]^{-2}\left[\sum_{j=1}^{m_k}(1-B_{kj})^2V(\bar{z}_{kj})\right.$$
$$\left.+\sum_{1\le j\ne j'\le m_k}(1-B_{kj})(1-B_{kj'})Cov(\bar{z}_{kj},z_{kj'})\right]$$
$$=\left[\sum_{j=1}^{m_k}(1-B_{kj})\right]^{-2}\left[\sum_{j=1}^{m_k}(1-B_{kj})^2\{\tau^2(1-B_{kj})^{-1}+\delta^2\}\right.$$
$$\left.+\sum_{1\le j\ne j'\le m_k}(1-B_{kj})(1-B_{kj'})\delta^2\right]$$
$$=\tau^2\left[\sum_{j=1}^{m_k}(1-B_{kj})\right]^{-1}+\delta^2$$
$$=\tau^2\left[\sum_{j=1}^{m_k}(1-B_{kj})\right]^{-1}\left[1+\lambda_1\lambda_2^{-1}\sum_{j=1}^{m_k}(1-B_{kj})\right]$$
$$=\tau^2\left[g_k\sum_{j=1}^{m_k}(1-B_{kj})\right]^{-1}=\tau^2c_k^{-1}. \tag{4.125}$$

□

Remark 5. For the one stratum case, since $\delta^2=0$, ν^* reduces to $\tilde{\nu}$ of Scott and Smith (1969). Then e^*_{EB} reduces to the Bayes estimator of Scott and Smith (1969). The latter, however, derived the same using a hierarchical Bayes approach, a topic to be discussed in the next chapter. Also, this agreement between the HB and the EB approach will be brought out in a more general context in the next chapter.

So far, the variance components σ^2, τ^2 and δ^2 are all assumed to be known. For most practical situations, however, these parameters are unknown, and need to be estimated from the data. To this end, in view of the lack of distributional assumptions, one uses the basic ANOVA technique which is described below. Let $\bar{z}_k=\sum_{j=1}^{m_k}n_{kj}\bar{z}_{kj}/\sum_{j=1}^{m_k}n_{kj}$ and $\bar{z}=\sum_{k=1}^{L}n_k\bar{z}_k/\sum_{k=1}^{L}n_k$. Define

$$SS_1=\sum_{k=1}^{L}\sum_{j=1}^{m_k}\sum_{\{i\in s,\ i\in psu\ j\ \text{in}\ str_k\}}(z_i-\bar{z}_{kj})^2; \tag{4.126}$$

$$SS_2 = \sum_{k=1}^{L} \sum_{j=1}^{m_k} n_{kj}(\bar{z}_{kj} - \bar{z}_k)^2; \tag{4.127}$$

$$SS_3 = \sum_{k=1}^{L} n_k(\bar{z}_k - \bar{z})^2. \tag{4.128}$$

Letting $n_T = \sum_{k=1}^{L} n_k$ and $m_T = \sum_{k=1}^{L} m_k$ and using Lemma 4.4.1 of the previous section, one gets

$$E(SS_1) = (n_T - m_T)\sigma^2; \tag{4.129}$$

$$E(SS_2) = \left(n_T - \sum_{k=1}^{L} n_k^{-1} \sum_{j=1}^{m_k} n_{kj}^2 \right) \tau^2 + (m_T - L)\sigma^2; \tag{4.130}$$

$$\begin{aligned} E(SS_3) &= \left(n_T - n_T^{-1} \sum_{k=1}^{L} n_k^2 \right) \delta^2 \\ &+ \left\{ \sum_{k=1}^{L} (n_k^{-1} - n_T^{-1}) \sum_{j=1}^{m_k} n_{kj}^2 \right\} + (L-1)\sigma^2. \end{aligned} \tag{4.131}$$

Define now

$$MS_1 = SS_1/(n_T - m_T)\sigma^2;$$
$$MS_2 = SS_2/(m_T - L);$$
$$MS_3 = SS_3/(L-1).$$

Then

$$E(MS_1) = \sigma^2. \tag{4.132}$$

Also, writing

$$h_1 = (m_T - L)^{-1} \left(n_T - n_T^{-1} \sum_{k=1}^{L} n_k^2 \right),$$

$$h_2 = (L-1)^{-1} \sum_{k=1}^{L} (n_k^{-1} - n_T^{-1}) \sum_{j=1}^{m_k} n_{kj}^2,$$

and

$$h_3 = (L-1)^{-1} \left(n_T - n_T^{-1} \sum_{k=1}^{L} n_k^2 \right),$$

$$E(MS_1) = \sigma^2;\ E(MS_2) = \sigma^2 + h_1\tau^2;\ E(MS_3) = \sigma^2 + h_2\tau^2 + h_3\delta^2. \tag{4.133}$$

This leads to the identities

$$E[(MS_2 - MS_1)h_1^{-1}] = \tau^2;$$

$$E[(MS_3 - MS_1 - h_2h_1^{-1}(MS_2 - MS_1)h_3^{-1}] = \delta^2.$$

Hence, σ^2, λ_1 and λ_2 are estimated respectively by

$$\sigma^2 = MS_1; \tag{4.134}$$

$$\hat{\lambda}_1 = max\{0, (MS_2/MS_1 - 1)^{-1}h_1\}; \tag{4.135}$$

$$\hat{\lambda}_1\hat{\lambda}_2^{-1} = max\{0, [h_1(MS_3 - MS_1)(MS_2 - MS_1)^{-1} - h_2]h_3^{-1}\}. \tag{4.136}$$

Next let

$$\hat{B}_{kj} = \hat{\lambda}_1(\hat{\lambda}_1 + n_{kj})^{-1};$$

$$\hat{g}_k = \left[1 + \hat{\lambda}_1\hat{\lambda}_2^{-1}\sum_{j=1}^{m_k}(1 - \hat{B}_{kj})\right]^{-1};$$

$$\hat{\nu} = \left[\sum_{k=1}^{L}\hat{g}_k\sum_{j=1}^{m_k}(1 - \hat{B}_{kj})\right]^{-1}\left[\sum_{k=1}^{L}\hat{g}_k\sum_{j=1}^{m_k}(1 - \hat{B}_{kj})\bar{z}_{kj}\right];$$

$$\hat{u}_k = \hat{g}_k\left(\hat{\nu} + \hat{\lambda}_1\hat{\lambda}_2^{-1}\sum_{j=1}^{m_k}(1 - \hat{B}_{kj})\bar{z}_{kj}\right).$$

An EB estimator of $\boldsymbol{\mu}$ is now given by $\boldsymbol{e}_{EB} = (e_{EB}^1, \ldots, e_{EB}^L)^T$, where

$$e_{EB}^k = \hat{u}_k + \sum_{j=1}^{m_k} r_{kj}(1 - f_{kj}\hat{B}_{kj})(\bar{z}_{kj} - \hat{u}_k). \tag{4.137}$$

We state below a result which establishes the asymptotic optimality of the EB estimator $\boldsymbol{e}_{EB}$. The proof is long and technical, and is omitted. One may refer to Lahiri (1986) for details.

Theorem 4.9 *Assume that*

(i) $inf_{1\le k\le L} inf_{1\le j\le m_k} n_{kj} \ge 1$;

(ii) $sup_{1\le k\le L} n_k = C < \infty$;

(iii) $inf_{1\le k\le L} M_k/N_k > 0$;

(iv) $E[\mu_4(\theta_{11})] < \infty$;

(v) $E[\mu_4(q_1)] < \infty$.

Then denoting the prior given in (A)–(C) by ξ, and the Bayes risk of an estimator e of μ under the loss given in (4.92) by $r(\xi; e, \mu)$, one gets

$$r(\xi; e_B, \mu) - r(\xi; e_{EB}, \mu) \to 0 \text{ as } k \to \infty.$$

We now see an application of the methods developed in this section for the analysis of some actual data. The data were collected with the primary objective of comparing the quality of radiation therapy of cancer patients among subpopulations of a population of facilities where radiation therapy was practised. In this case the finite population of units is actually the sampling units arising from a 1978 survey of patients suffering from cervical cancer. For conducting this survey, radiation therapy facilities were grouped into several strata that were thought to be relatively homogeneous in the quality of care that patients received. The five strata considered here correspond to strata 1, 2, 4, 5 and 6 of Calvin and Sedransk (1991). The number of facilities contained in these five strata are 10, 15, 11, 30, and 11 respectively, and are treated as primary sampling units (psu's). Among these psu's, a 1/3 simple random sampling is used resulting in the selection of 3, 5, 4, 10, and 4 psu's from the five strata. From each selected psu with p patient records, a simple random sample of size $[(p+1)/2]$ is selected where $[u]$ denotes the integer part of u.

The present analysis considers 'pretreatment' scores for each patient. For a given patient, for each disease site, a committee of experts identified a set of services and procedures (S/P's) that were thought to be of prime importance for a complete pretreatment evaluation and for planning and monitoring therapy. The committee also assigned weights (0.5 to 4.0) to these S/P's to indicate their relative importance. Then, for each patient, a score is defined by $\sum_i W_i^* Z_i / \sum_i W_i^*$, where W_i^* is the corresponding weight. The larger the score, the closer the patient's care conforms to acceptable standards of care.

Let y_i denote the score for the ith patient. Although the y_i's lie between 0 and 1, these are weighted averages of independent Bernoulli variables, and a normal approximation due to the CLT is not totally out of the way.

We compare the EB estimators derived in this section with four other estimators. These estimators are given in the following four

equations

$$e_U^k = (M_k/m_k)\left(\sum_{j=1}^{m_k} N_{kj}\bar{z}_{kj}\right)/\left(\sum_{j=1}^{m_k} N_{kj}\right) \tag{4.138}$$

(a design unbiased estimator);

$$\begin{aligned} e_R^k &= \left(\sum_{j=1}^{M_k} N_{kj}\right)^{-1}\left[\sum_{i\in s,\ i\in psu\ j\ \text{in}\ str_k} z_i + \sum_{j=1}^{m_k}(N_{kj}-n_{kj})\bar{z}_{kj}\right. \\ &+ \left.\left(\sum_{j=1}^{m_k} N_{kj}\bar{z}_{kj}/\sum_{j=1}^{m_k} N_{kj}\right)\left(\sum_{j=m_k+1}^{M_k} N_{kj}\right)\right] \end{aligned} \tag{4.139}$$

(the ratio-type estimator);

$$\begin{aligned} e_0^k &= \left(\sum_{j=1}^{M_k} N_{kj}\right)^{-1}\left[\sum_{i\in s,\ i\in psu\ j\ \text{in}\ str_k} z_i\right. \\ &+ \left.\left(\sum_{j=1}^{M_k} N_{kj} - \sum_{j=1}^{m_k} n_{kj}\right)\left(\sum_{j=1}^{m_k} n_{kj}\bar{z}_{kj}\right)/\sum_{j=1}^{m_k} n_{kj}\right] \end{aligned} \tag{4.140}$$

(the expansion estimator);

$$\begin{aligned} e_{RO}^k &= \left(\sum_{j=1}^{M_k} N_{kj}\right)^{-1}\left[\sum_{i\in s,\ i\in psu\ j\ \text{in}\ str_k} z_i\right. \\ &+ \left.\sum_{j=1}^{m_k}(N_{kj}-n_{kj})\bar{z}_{kj} + \left(\sum_{j=1}^{m_k}\bar{z}_{kj}/m_k\right)\sum_{j=m_k+1}^{M_k} N_{kj}\right]. \end{aligned} \tag{4.141}$$

(Royall's estimator).

The estimators e_R^k, e_0^k and e_{RO}^k are all based on the predicted values of the unobserved units on the basis of the sampled units. However, in contrast to the present HB model, they can possibly be justified on the basis of some other models as given for example in Royall (1976). Table 4.5 provides the true population means as well as the five different estimates for each stratum.

We found the average absolute biases of the EB estimates, the design unbiased estimates, the ratio-type estimates, the expansion estimates, and Royall's estimates for the given dataset to be respectively by 0.0316, 0.1293, 0.0628, 0.0601, 0.0484. Thus the EB

Table 4.5 *The true means μ_k and their estimates.*

k	μ_k	e^k_{EB}	e^k_U	e^k_R	e^k_0	e^k_{RO}
1	0.7333	0.8031	0.7120	0.9185	0.9219	0.9321
2	0.7615	0.7644	0.9100	0.7721	0.7682	0.7504
3	0.7448	0.7684	0.7821	0.7838	0.7830	0.7504
4	0.6893	0.7497	0.8965	0.7386	0.7400	0.7153
5	0.7455	0.7418	0.9806	0.7131	0.7265	0.7200

estimates have much greater edge over the other four estimates in terms of absolute bias. Also, the total sum of squared deviations of the EB estimates from the true means is 0.0091. The corresponding figures for e_U, e_R, e_0 and e_{RO} turn out to be 0.1211, 0.0391 0.0400 and 0.0409 respectively. Hence, the percentage reduction in the total sum of squared deviations for the EB estimates is 92.5, 76.7, 77.3 and 77.8 in comparison with the design unbiased estimates, the ratio-type estimates, the expansion estimates and Royall's estimates respectively.

The improvement of the EB estimator over the other four estimators is indeed startling. One possible explanation of this fact is that many of the other estimators are optimal under models which do not take into account variation in the primary sampling units. The Bayesian model takes into account this extra source of variation, and thus produces more reliable estimators.

4.6 Auxiliary information

In the previous section, we considered robust EB estimation of stratum means when no auxiliary information was available. In actual surveys, however, auxiliary information exists, judicious use of which can lead to estimators with enhanced precision when compared to the direct survey estimators. For instance, in estimating the per capita income for small places, Fay and Herriot (1979) used as auxiliary variables the corresponding county averages, tax-return data for 1969 and data on housing from the 1970 census. EB estimators constructed by these authors which incorporated the aforementioned auxiliary information were far superior to the survey estimators in the sense that the estimated standard errors

and coefficients of variation associated with the former were much smaller in comparison with the latter. For estimating the acreage of corn, and the acreage of soybean for 12 counties in North Central Iowa, Battese *et al.* (1988) used auxiliary information from the Landsat Satellite data in addition to the data available from the USDA. on the acreage of these crops from the 36 sampling units.

We shall develop in this section a general EB methodology for estimation of stratum means in the presence of auxiliary information without making any distributional assumptions, but invoking as before the notion of posterior linearity. The following model which is a slight extension of the models given in the preceding two sections is used.

(I) Conditional on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^T, they_i$ are mutually independent, and for the units belonging to str_k we have that $E(y_i|\boldsymbol{\theta}) = \theta_k$ and $V(y_i|\boldsymbol{\theta}) = \mu_2(\theta_k), k = 1, \ldots, L$;

(II) θ_k are independent with $E(\theta_k) = \boldsymbol{x}_k^T\boldsymbol{\beta}, V(\theta_k) = \tau^2$, where the $\boldsymbol{x}_k$ $(p \times 1)$ are the known design vectors, and $\boldsymbol{\beta}$ $(p \times 1)$ is the unknown regression vector. It is assumed that $p < L$;

(III) $0 < \sigma^2 < \infty$.

Note that in the previous section, $p = 1, \boldsymbol{x}_k = 1$ for all k and $\beta_1 = \mu$. As in the previous section, we assume posterior linearity, namely

$$E(\theta_k|z) = \sum_{i \in str_k} a_{ki} z_i + b_k, \; k = 1, \ldots, L, \tag{4.142}$$

where the a_{ki} and b_k are constants not depending on the y_i. Once again, using the result of Goldstein (1975), it follows that

$$E(\theta_k|z) = (1 - B_k)\bar{z}_k + B_k \boldsymbol{x}_k^T \boldsymbol{\beta}, \tag{4.143}$$

where, as before, $B_k = \lambda/(\lambda + n_k)$ and $\lambda = \sigma^2/\tau^2$. Then the Bayes estimator of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^T$ is $\hat{\boldsymbol{\mu}}^B = (\hat{\mu}_1^B, \ldots, \hat{\mu}_L^B)^T$, where

$$\hat{\mu}_k^B = \bar{z}_k - f_k B_k (\bar{z}_k - \boldsymbol{x}_k^T \boldsymbol{\beta}), \tag{4.144}$$

where $f_k = (N_k - n_k)/N_k$ denotes once again the finite population correction factor for stratum k.

In an EB analysis σ^2, τ^2 and $\boldsymbol{\beta}$ are all unknown, and need to be estimated from the data. However, as before, it is convenient first to assume that σ^2 and τ^2 (and hence λ) are known, but $\boldsymbol{\beta}$ is unknown. Marginally,

$$E(\bar{z}_k) = EE(\bar{z}_k|\boldsymbol{\theta}) = E(\theta_k) = \boldsymbol{x}_k^T \boldsymbol{\beta}; \tag{4.145}$$

$$V(\bar{z}_k) = \sigma^2\lambda^{-1}(1 - B_k)^{-1}. \tag{4.146}$$

Write $\boldsymbol{X}^T = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$ and assume that $\text{rank}(\boldsymbol{X}) = p$. Then the BLUE of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{D}\tilde{\boldsymbol{\mu}}_0, \tag{4.147}$$

where $\boldsymbol{D} = diag(1 - B_1, \ldots, 1 - B_L)$ and $\tilde{\boldsymbol{\mu}}_0 = (\bar{z}_1, \ldots, \bar{z}_L)^T$. The resulting EB estimator of $\boldsymbol{\mu}$ is $\tilde{\boldsymbol{\mu}}^{EB} = (\tilde{\mu}_1^{EB}, \ldots, \tilde{\mu}_L^{EB})^T$, where

$$\tilde{\mu}_k^{EB} = \bar{z}_k - f_k B_k(\bar{z}_k - \boldsymbol{x}_k^T\tilde{\boldsymbol{\beta}}). \tag{4.148}$$

We compare the performance of $\tilde{\boldsymbol{\mu}}^{EB}$ with the classical estimator $\tilde{\boldsymbol{\mu}}_0$ of $\boldsymbol{\mu}$ in terms of their Bayes risks. Denote the prior given in (I)–(III) by ξ. Once again, let $r(\xi, \boldsymbol{e})$ denote the Bayes risk of an estimator $\boldsymbol{e}$ of $\boldsymbol{\mu}$ with respect to the prior ξ and the loss given in (4.22) on page 168. In this case,one gets

$$\begin{aligned} r(\xi, \tilde{\boldsymbol{\mu}}^0) - r(\xi, \tilde{\boldsymbol{\mu}}^B) &= \sum_{k=1}^{L} f_k^2 B_k^2 E(\bar{z}_k - \boldsymbol{x}_k^T\boldsymbol{\beta})^2/L \\ &= \sigma^2\lambda^{-1}L^{-1}\sum_{k=1}^{L} f_k^2 B_k^2(1 - B_k)^{-1}; \end{aligned} \tag{4.149}$$

$$\begin{aligned} r(\xi, \tilde{\boldsymbol{\mu}}^{EB}) - r(\xi, \tilde{\boldsymbol{\mu}}^B) &= \sum_{k=1}^{L} f_k^2 B_k^2 E(\boldsymbol{x}_k^T\tilde{\boldsymbol{\beta}} - \boldsymbol{x}_k^T\boldsymbol{\beta})^2/L \\ &= \sigma^2\lambda^{-1}L^{-1}\sum_{k=1}^{L} f_k^2 B_k^2 \boldsymbol{x}_k^T(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}. \end{aligned} \tag{4.150}$$

This leads to the relative savings loss of $\tilde{\boldsymbol{\mu}}^{EB}$ with respect to $\tilde{\boldsymbol{\mu}}_0$ as

$$\begin{aligned} RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0) &= \sum_k f_k^2 B_k^2 \boldsymbol{x}_k^T(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_k \\ &\div \sum_k f_k^2 B_k^2(1 - B_k)^{-1}. \end{aligned} \tag{4.151}$$

The first theorem of this section provides the asymptotic (as $L \to \infty$) behavior of $RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0)$.

Theorem 4.10 *Under the assumptions of Theorem 4.4*

$$RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0) = O(L^{-1}) \text{ as } L \to \infty. \tag{4.152}$$

Proof. First note that since $1 - B_k = n_k/(\lambda + n_k) \geq 1/(\lambda + 1)$,

$$\begin{aligned} \boldsymbol{x}_k^T(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_k &= \boldsymbol{x}_k^T\left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{x}_k \\ &\leq (\lambda+1)^{-1}\boldsymbol{x}_k^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_k. \end{aligned} \tag{4.153}$$

Also, $f_k^2 B_k^2 \leq 1$ for all $k = 1, \ldots, L$. Hence, from (4.150),

$$\begin{aligned} &[(\lambda^2+\lambda)[r(\xi, \tilde{\boldsymbol{\mu}}^{EB}) - r(\xi, \tilde{\boldsymbol{\mu}}^{B})] \\ &\leq \sigma^2 L^{-1}\sum_{k=1}^{L}\boldsymbol{x}_k^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_k \\ &= \sigma^2 L^{-1} tr\left(\sum_{k=1}^{L}\boldsymbol{x}_k^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_k\right) \\ &= \sigma^2 L^{-1} tr\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\sum_{k=1}^{L}\boldsymbol{x}_k\boldsymbol{x}_k^T\right) \\ &= \sigma^2 L^{-1} tr(\boldsymbol{I}_p) \\ &= \sigma^2 L^{-1} p. \end{aligned} \tag{4.154}$$

On the other hand, since $B_k = \lambda(\lambda + n_k)^{-1} \geq \lambda(\lambda + C)^{-1}$ and $f_k = 1 - n_k N_k^{-1} \geq 1 - n_k(n_k+1)^{-1} = (n_k+1)^{-1} \geq (C+1)^{-1}$, it follows from (4.149) that

$$r(\xi, \tilde{\boldsymbol{\mu}}_0) \geq (C+1)^{-2}\lambda^2(\lambda+C)^{-2}(\lambda+1)^{-1}. \tag{4.155}$$

Combining (4.151), (4.154) and (4.155), it follows that

$$RSL(\xi; \tilde{\boldsymbol{\mu}}^{EB}, \tilde{\boldsymbol{\mu}}_0) = O(L^{-1}). \tag{4.156}$$

□

Remark 1. Note that it follows from (4.154) that $\tilde{\boldsymbol{\mu}}^{EB}$ is asymptotically optimal in the sense of Robbins (1956).

Next consider the situation when $\boldsymbol{\beta}, \sigma^2$ and τ^2 are all unknown, and need to be estimated from the data. Write $n_T = \sum_{k=1}^{L}$. It is assumed that $n_T \geq 2$.

The estimation proceeds as in the previous section. Define SSW and MSW as in the previous two sections, and as before, $E(MSW) = \sigma^2$. However, the estimation of τ^2 is much more complex, and one proceeds as in Carter and Rolf (1974) or Fay and Herriot (1979).

With this end, first note that

$$E(\bar{z}_k - \boldsymbol{x}_k^T\boldsymbol{\beta})^2 = V(\bar{z}_k) + V(\boldsymbol{x}_k^T\tilde{\boldsymbol{\beta}}) - 2Cov(\bar{z}_k, \boldsymbol{x}_k^T\tilde{\boldsymbol{\beta}}) = \sigma^2\lambda^{-1}[(1-B_k)^{-1} - \boldsymbol{x}_k^T(\boldsymbol{X}^T\boldsymbol{D}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_k]. \tag{4.157}$$

This implies

$$E\left[\sum_{k=1}^{L}(1-B_k)(\bar{z}_k - \boldsymbol{x}_k^T\tilde{\boldsymbol{\beta}})^2\right] = \sigma^2\lambda^{-1}\left[L - \sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k^T(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T)^{-1}\boldsymbol{x}_k\right]. \tag{4.158}$$

Next simplify

$$\begin{aligned}
&\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k^T\left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{x}_k \\
&\quad = tr\left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k^T(\sum_{k=1}^{L}(1-B_K)\boldsymbol{x}_k\boldsymbol{x}_k^T)^{-1}\boldsymbol{x}_k\right) \\
&\quad = tr\left(\left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right) \\
&\quad = tr(\boldsymbol{I}_p) = p.
\end{aligned} \tag{4.159}$$

Then

$$E\left[\sum_{k=1}^{L}(1-B_k)(\bar{z}_k - \boldsymbol{x}_k^T\tilde{\boldsymbol{\beta}})^2\right] = (L-p)\sigma^2\lambda^{-1} = (L-p)\tau^2. \tag{4.160}$$

Now, using the method of moments τ^2 is estimated by solving iteratively equation (4.147) and

$$\sum_{k=1}^{L}(1-B_k)(\bar{z}_k - \boldsymbol{x}_k^T\tilde{\boldsymbol{\beta}})^2/(L-p) = \tau^2. \tag{4.161}$$

Now $\hat{\lambda} = \hat{\sigma}^2/\hat{\tau}^2$ and $\hat{B} = \hat{\lambda}/(\hat{\lambda} + n_k)$. Also, let $\hat{\boldsymbol{D}} = diag(1 - \hat{B}_1, \ldots, 1 - \hat{B}_k)$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{D}}\tilde{\mu}_0$. Substituting all these estimators, the final EB estimator of $\boldsymbol{\mu}$ turns out to be $\hat{\boldsymbol{\mu}}^{EB} =$

$(\hat{\mu}_1^{EB}, \ldots, \hat{\mu}_L^{EB})^T$ where

$$\hat{\mu}_k^{EB} = \bar{z}_k - f_k \hat{B}_k (\bar{z}_k - \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}). \tag{4.162}$$

The asymptotic optimality of this estimator can be proved by imposing certain conditions on the design matrix $\boldsymbol{X}$. The details are omitted for brevity.

Next, we consider the case when only some summary statistics are available from each stratum, and introduce a slight variation of the model given in (I)–(III). More specifically, the following model is assumed :

(i) conditional on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^T$, z_k are independent $N(\theta_k, V_k)$, where V_k (> 0) are known constants.

(ii) θ_k are independent $N(\boldsymbol{x}_k^T \boldsymbol{\beta}, \tau^2)$.

Suppose also that instead of the finite population means, one is interested in inference about the θ_k themselves. The posterior mean of θ_k is given by

$$E(\theta_k | \boldsymbol{z}) = (1 - B_k) z_k + B_k \boldsymbol{x}_k^T \boldsymbol{\beta}, \tag{4.163}$$

where $B_k = V_k/(V_k + \tau^2)$. Clearly, the normality assumption can be replaced by the assumption of posterior linearity.

Once again for an EB analysis, $\boldsymbol{\beta}$ and τ^2 are unknown, and are estimated from the data. These estimators are obtained by solving iteratively equations similar to (4.147) and (4.162) with z_k replacing $\bar{z}_k$ in the definition of $\tilde{\mu}_0$, and B_k being defined as in the preceding paragraph. The EB estimator of θ_k is given by

$$\hat{\theta}_k = (1 - \hat{B}_k) z_k + \hat{B}_k \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}. \tag{4.164}$$

The corresponding EB estimator of μ_k is given by

$$\hat{\mu}_k^{EB} = (1 - f_k \hat{B}_k) z_k + f_k \hat{B}_k \boldsymbol{x}_k^T \hat{\boldsymbol{\beta}}. \tag{4.165}$$

We shall see now an application of the above methodology in a real-life problem. This relates to estimation of median income of four-person families. The US department of health and human services (HHS) administers a programme of energy assistance to low-income families. Eligibility for the programme is determined by a formula where the most important variable is an estimate of the current median income for four-person families by states (the 50 states and the District of Columbia).

The Bureau of the Census, by an informal agreement, has provided such estimates to the HHS by a linear regression methodology since the latter part of the 1970s. In its original approach

which was continued up to the mid-1980s, the sample estimates Y of the state medians for the most current year c, as well as the associated standard errors, are first obtained from the Current Population Survey (CPS). These estimates Y were used as dependent variables in a linear regression procedure with the single predictor variable

$$\begin{aligned} \textit{Adjusted census median}(c) &= [\textit{BEA PCI}(c)/\textit{BEA PCI}(b)] \\ &\times \textit{census median}(b). \end{aligned}$$

In the above, census median(b) represents the median income of four-person families in the state for the base year b from the most recently available decennial census. BEA PCI(c) and BEA PCI(b) represent, respectively, estimates of per capita income produced by the Bureau of Economic Analysis of the Department of Commerce for the current year c and the base year b. It may be pointed out also that for each decennial census year, the household income figures are collected for the preceding calendar year. For example, the 1980 census collected 1979 income figures from the different households. Thus, the adjusted census median is arrived at by multiplying the base year median by the proportional growth in the per capita income. The linear regression that was used is

$$Y_i = \beta_0 + \beta_1 x_i + e_i,\ i = 1, \ldots, 51,$$

where Y_i denotes the median income of four-person families for the ith state, and x_i denotes the corresponding adjusted census median.

Next a weighted average of the CPS sample estimate of the current median income and the regression estimate is obtained, weighting the CPS estimate inversely proportional to the sampling variance, and the regression estimate inversely proportional to the residual mean square due to lack of fit for the 1969 median values obtained from the 1970 census by the model with 1959 used as the base year. Finally, the above composite estimate was constrained not to deviate by more than one standard deviation from the corresponding CPS sample estimate.

Fay (1987) suggested another linear regression model which includes the base year census median b in addition to the adjusted census median c. The idea is to adjust for any possible overstatement of the effect of change in BEA income upon the median income of four-person families. The alternative model is given by (i) and (ii) of this section, with $x_i = (1, x_{1i}, x_{2i})^T$; x_{1i} and x_{2i} repre-

sent respectively the adjusted census median income, and the base year census median income for the ith state.

We provide next the EB estimates for 1979 using methods of this section. These estimates are compared against the CPS estimates and the Bureau of the Census estimates treating the 1980 census figures as the 'truth'. Tables 4.6 and 4.7 provide the 'truth' as well as all the estimates. Due to smallness of the sample size relative to the population sample size, the finite population correction can be ignored in this case. In other words $f_k = 1$ for all k.

Next, these estimates are compared according to several criteria. Suppose e_{iTR} denotes the true median income for the ith state, and e_i denotes any estimate of e_{iTR}. Then one computes

Average relative bias $= (51)^{-1} \sum_{i=1}^{51} |e_i - e_{iTR}| e_{iTR}^{-1}$;

Average squared relative bias $= (51)^{-1} \sum_{i=1}^{51} [e_i - e_{iTR}]^2 e_{iTR}^{-2}$;

Average absolute bias $= (51)^{-1} \sum_{i=1}^{51} |e_i - e_{iTR}|$;

Average squared deviation $= (51)^{-1} \sum_{i=1}^{51} [e_i - e_{iTR}]^2$.

Under all the four criteria, the CPS estimate has the worst performance, while the EB estimate performs the best. In terms of the average relative bias, the EB estimate improves on the CPS estimate by 59%, while it improves on the Bureau estimate by 37.1%. Under the criterion of squared relative bias, the corresponding improvements are 80.0% and 58.8%. Similarly, for average absolute bias as the criterion, the improvements are 58.7% and 37.7%. Finally, using average squared deviations as the criterion, the respective percentage improvements are 79.5% and 60.0%.

4.7 Nested error regression models

As an alternative to the model considered in the previous section, we consider a model where in addition to the area-specific random effects, there exist unit-specific design vectors. Such models are used by Battese *et al.* (1988), and Datta and Ghosh (1991) for estimating the area under corn and soybean for 12 counties in North Central Iowa. Under the present model, if unit i belongs to stratum k, we model the corresponding y_i as $y_i = \boldsymbol{x}_{ki}^T \boldsymbol{b} + v_k + e_i$, where certain assumptions about the v_k and the e_i will be made later. It is assumed that the population size of the kth stratum is N_k, and a sample of size n_k is drawn from the kth stratum. Write $\boldsymbol{y}_k^P$ $(N_k \times 1)$ as the vector of characteristics from the kth stratum

Table 4.6 *Median income estimates for four-person families for states 1–34 in 1979 (in dollars).*

State	Truth	Bureau est.	CPS est.	EB est.
1	18319	18074	18084	18818
2	22027	22335	23129	22468
3	19424	19314	18438	20535
4	23772	23786	25324	23752
5	22107	21636	23597	21744
6	25712	24410	25037	25447
7	22669	21082	21852	22614
8	26014	24640	25486	25133
9	22266	22314	23326	21937
10	23279	22528	22397	23022
11	23014	22614	22716	22623
12	25410	24265	24294	24528
13	25111	24422	24103	24471
14	23320	23518	23270	23986
15	24044	24409	25137	24193
16	22351	22567	23851	22338
17	21891	21294	20743	21745
18	20511	19520	18567	20126
19	18674	19209	19826	19476
20	21438	20749	22603	20689
21	22127	22848	22014	22360
22	23627	21184	21860	21800
23	26203	24686	26235	24896
24	21862	21310	20232	22210
25	22757	22976	24160	22600
26	20214	18876	18274	19376
27	19772	19648	20296	19522
28	19944	20154	19282	19588
29	20668	21574	22687	20550
30	21086	20757	19675	21534
31	19685	19138	18657	19245
32	19693	19437	19776	19221
33	19926	18613	17978	19816
34	18150	17672	19167	17841

Table 4.7 *Median income estimates for four-person families for states 35–51 in 1979 (in dollars).*

State	Truth	Bureau est.	CPS est.	EB est.
35	17893	18493	18917	18056
36	21412	20166	18965	20207
37	20659	20852	19295	20887
38	22521	23416	22963	22355
39	20776	20051	21216	20116
40	19961	20429	21533	20259
41	24641	22673	22013	24233
42	23757	25228	26746	23842
43	19257	21032	19968	20114
44	21924	23000	21733	22631
45	21572	21250	20727	21327
46	24438	25457	25447	24215
47	24394	24410	25463	24173
48	22688	24031	24251	23227
49	24752	25109	24265	24988
50	31018	31037	30788	30139
51	24966	24582	23744	24992

in the population, and write the model as

$$\boldsymbol{y}_k^P = \boldsymbol{X}_k^P \boldsymbol{b} + v_k \mathbf{1}_{N_k} + \boldsymbol{e}_k^P, \; k = 1, \ldots, L. \tag{4.166}$$

Such models are usually referred to as **nested error regression models**. We partition $\boldsymbol{y}_k^P$ as $\boldsymbol{y}_k^P = (\boldsymbol{z}_k^T, \boldsymbol{y}_k^T(\bar{s}))^T$, where $\boldsymbol{z}_k$ denotes the vector of observed units from the ith stratum, and $\boldsymbol{y}_k(\bar{s})$ denotes the vector of unobserved units from the ith stratum. Correspondingly, partition $\boldsymbol{X}_k^P$ as $\boldsymbol{X}_k^P = (\boldsymbol{X}_k^T(s), \boldsymbol{X}_k^T(\bar{s}))^T$, and partition $\boldsymbol{e}_k^P$ into $\boldsymbol{e}_k^P = (\boldsymbol{e}_k^T(s), \boldsymbol{e}_k^T(\bar{s}))^T$. This leads to the representation

$$\begin{aligned} \boldsymbol{z}_k &= \boldsymbol{X}_k(s)\boldsymbol{b} + v_k \mathbf{1}_{n_k} + \boldsymbol{e}_k(s); \\ \boldsymbol{y}_k(\bar{s}) &= \boldsymbol{X}_k(\bar{s})\boldsymbol{b} + v_k \mathbf{1}_{N_k - n_k} + \boldsymbol{e}_k(\bar{s}), \end{aligned} \tag{4.167}$$

for all $k = 1, \ldots, L$. It is assumed that $(v_1, \boldsymbol{e}_1^P), \ldots, (v_L, \boldsymbol{e}_L^P)$ are mutually independent, and v_k and $\boldsymbol{e}_k$ are also independently distributed with $v_k \sim N(0, \sigma_v^2 c_k)$, and $\boldsymbol{e}_k \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{D}_k)$. In the above, $c_k > 0$, and $\boldsymbol{D}_k$ is a diagonal matrix of order N_k with

each diagonal element positive. The objective is once again to find the conditional distribution of $\boldsymbol{y}_k(\bar{s})$ given $\boldsymbol{z}$, and find posterior means, variances and covariances of quantities of interest. More particularly, we will be interested in finding posterior means, variances and covariances of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)^T$, where as before $\mu_k = N_k^{-1} \sum_{i \in str_k} y_i$, $k = 1, \ldots, L$.

Clearly, (4.167) can be regarded as a general mixed-effects model. A full Bayesian analysis will assign distributions to $\boldsymbol{b}$, σ_v^2 and σ^2, and will be considered in the next chapter. In this section, however, we consider an EB approach for solving this problem which treats $\boldsymbol{b}$, σ_v^2 and σ^2 as parameters to be estimated from the data.

First, rewrite the model given in (4.167) to get the joint distribution of

$$\begin{bmatrix} \boldsymbol{z}_k \\ \boldsymbol{y}_k(\bar{s}) \end{bmatrix} \tag{4.168}$$

as multivariate normal with mean vector

$$\begin{bmatrix} \boldsymbol{X}_k(s)\boldsymbol{b} \\ \boldsymbol{X}_k(\bar{s})\boldsymbol{b} \end{bmatrix},$$

and covariance matrix

$$\begin{bmatrix} \sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k} & \sigma_v^2 c_k \boldsymbol{J}_{n_k, N_k - n_k} \\ \sigma_v^2 c_k \boldsymbol{J}_{N_k - n_k, n_k} & \sigma^2 \boldsymbol{D}_k(\bar{s}) + \sigma_v^2 c_k \boldsymbol{J}_{N_k - n_k} \end{bmatrix},$$

where $\boldsymbol{J}_{u,v} = \mathbf{1}_u \mathbf{1}_v^T$ and $\boldsymbol{J}_u = \boldsymbol{J}_{u,u}$. Also, the diagonal matrix $\boldsymbol{D}_k$ is partitioned into $\boldsymbol{D}_k = Diag(\boldsymbol{D}_k(s), \boldsymbol{D}_k(\bar{s}))$ where $\boldsymbol{D}_k(s) = V(\boldsymbol{z}_k)$ and $\boldsymbol{D}_k(\bar{s}) = V(\boldsymbol{y}_k(\bar{s}))$.

Based on (4.168) one gets the posterior pdf of $\boldsymbol{y}_k(\bar{s})$ given $\boldsymbol{z}$ as normal with mean

$$\boldsymbol{X}_k(\bar{s})\boldsymbol{b} + \sigma_v^2 c_k \boldsymbol{J}_{N_k - n_k, n_k} (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\boldsymbol{b})$$

and variance

$$\begin{aligned} \sigma^2 \boldsymbol{D}_k(\bar{s}) \quad &+ \quad \sigma_v^2 c_k \boldsymbol{J}_{N_k - n_k} - \sigma_v^4 c_k^2 \boldsymbol{J}_{N_k - n_k, n_k} \\ &\times \quad (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} \boldsymbol{J}_{n_k, N_k - n_k}. \end{aligned} \tag{4.169}$$

Write

$$h_k = \mathbf{1}_{n_k}^T \boldsymbol{D}_k^{-1}(s) \mathbf{1}_{n_k} \; \gamma_k = \sigma^2 (\sigma^2 + \sigma_v^2 c_k h_k)^{-1}. \tag{4.170}$$

Then, using Exercise 2.8 of Rao (1973),

$$\begin{aligned} &(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} \\ &\quad = \sigma^{-2} [\boldsymbol{D}_k^{-1}(s)] \end{aligned}$$

$$-h_k^{-1}(1-\gamma_k)\boldsymbol{D}_k^{-1}(s)\boldsymbol{J}_{n_k}\boldsymbol{D}_k^{-1}(s)]. \tag{4.171}$$

Hence,

$$\begin{aligned}
&\sigma_v^2 c_k \boldsymbol{J}_{N_k-n_k,n_k}(\sigma^2\boldsymbol{D}_k(s)+\sigma_v^2 c_k\boldsymbol{J}_{n_k})^{-1}| \\
&\quad = \sigma_v^2 c_k \mathbf{1}_{N_k-n_k}\sigma^{-2}\mathbf{1}_{n_k}^T(1-(1-\gamma_k))\boldsymbol{D}_k^{-1}(s) \\
&\quad = h_k^{-1}(1-\gamma_k)\boldsymbol{J}_{N_k-n_k,n_k}\boldsymbol{D}_k^{-1}(s).
\end{aligned} \tag{4.172}$$

Hence, writing

$$\bar{z}_{kw}(s) = \mathbf{1}_{n_k}^T\boldsymbol{D}_k^{-1}(s)\boldsymbol{z}_k/h_k,$$

and

$$\bar{\boldsymbol{x}}_{kw}^T(s) = \mathbf{1}_{n_k}^T\boldsymbol{D}_k^{-1}(s)\boldsymbol{X}_k(s)/h_k,$$

one gets

$$\begin{aligned}
E[\boldsymbol{y}_k(\bar{s})|\boldsymbol{z}] &= \boldsymbol{X}_k(\bar{s})\boldsymbol{b} \\
&+ h_k^{-1}(1-\gamma_k)\boldsymbol{J}_{N_k-n_k}\boldsymbol{D}_k^{-1}(s)(\boldsymbol{z}_k-\boldsymbol{X}_k(s)) \\
&= \boldsymbol{X}_k(\bar{s})\boldsymbol{b}+(1-\gamma_k)(\bar{z}_{kw}(s)-\bar{\boldsymbol{x}}_{kw}^T(s)\boldsymbol{b})\mathbf{1}_{N_k-n_k}.
\end{aligned} \tag{4.173}$$

Moreover,

$$\begin{aligned}
&\sigma^2\boldsymbol{D}_k(\bar{s})+\sigma_v^2 c_k\boldsymbol{J}_{N_k-n_k} \\
&\quad -\sigma_v^4 c_k^2\boldsymbol{J}_{N_k-n_k,n_k}(\sigma^2\boldsymbol{D}_k(s)+\sigma_v^2 c_k\boldsymbol{J}_{n_k})^{-1}\boldsymbol{J}_{n_k,N_k-n_k} \\
&\quad = \sigma^2\boldsymbol{D}_k(\bar{s})+\sigma_v^2 c_k\boldsymbol{J}_{N_k-n_k}-\sigma_v^2 c_k(1-\gamma_k)\boldsymbol{J}_{N_k-n_k} \\
&\quad = \sigma^2\boldsymbol{D}_k(\bar{s})+\sigma_v^2\gamma_k c_k\boldsymbol{J}_{N_k-n_k}.
\end{aligned} \tag{4.174}$$

This leads to the posterior distribution of $\boldsymbol{y}_k(\bar{s})$ given $\boldsymbol{z}$ as

$$\begin{aligned}
&N\Big[\boldsymbol{X}_k(\bar{s})\boldsymbol{b}+(1-\gamma_k)(\bar{z}_{kw}(s)-\bar{\boldsymbol{x}}_{kw}^T(s)\boldsymbol{b})\mathbf{1}_{N_k-n_k}, \\
&\sigma^2\Big(\boldsymbol{D}_k(\bar{s})+h_k^{-1}(1-\gamma_k)\boldsymbol{J}_{N_k-n_k}\Big)\Big].
\end{aligned} \tag{4.175}$$

Writing $\bar{y}_k(\bar{s}) = (N_k-n_k)^{-1}\mathbf{1}_{N_k-n_k}^T\boldsymbol{y}_k(\bar{s})$, one gets the posterior distribution of $\bar{y}_k(\bar{s})$ given $\boldsymbol{z}$ as

$$\begin{aligned}
&N\Big[\bar{\boldsymbol{x}}_k^T(\bar{s})\boldsymbol{b}+(1-\gamma_k)(\bar{z}_{kw}(s)-\bar{\boldsymbol{x}}_{kw}^T(s)\boldsymbol{b}), \\
&\sigma^2\Big((N_k-n_k)^{-2}\sum_{i\in str_k, i\in s'} d_{ki}+h_k^{-1}(1-\gamma_k)\Big)\Big],
\end{aligned} \tag{4.176}$$

where d_{ki}, $i = 1, \ldots, N_k$ denote the diagonal elements of $\boldsymbol{D}_k$, and $\bar{\boldsymbol{x}}_k(\bar{s}) = (N_k - n_k)^{-1} \sum_{i \in str_k, i \in \bar{s}} \boldsymbol{x}_{ki}$.

In an EB analysis, some or all of the parameters $\boldsymbol{b}$, σ_v^2 and σ^2 are unknown, and need to be estimated from the data. First assume that σ_v^2 and σ^2 are known, but $\boldsymbol{b}$ is unknown. Using the fact that $\boldsymbol{z}_k \sim N(\boldsymbol{X}_k(s)\boldsymbol{b}, \sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})$, the MLE or the best unbiased estimator of $\boldsymbol{b}$ is obtained by minimizing $\sum_{k=1}^{L} (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\boldsymbol{b})^T (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\boldsymbol{b})$ with respect to $\boldsymbol{b}$. Using the matrix result given in (4.170), this amounts to minimization of

$$\boldsymbol{b}^T \left[\sum_{k=1}^{L} \boldsymbol{X}_k^T(s)(\boldsymbol{D}_k^{-1}(s) - h_k^{-1}(1-\gamma_k)\boldsymbol{D}_k^{-1}(s)\boldsymbol{J}_{n_k}\boldsymbol{D}_k^{-1}(s))\boldsymbol{X}_k(s) \right] \boldsymbol{b}$$

$$-2\boldsymbol{b}^T \sum_{k=1}^{L} \boldsymbol{X}_k^T(s)(\boldsymbol{D}_k^{-1}(s) - h_k^{-1}(1-\gamma_k)\boldsymbol{D}_k^{-1}(s)\boldsymbol{J}_{n_k}\boldsymbol{D}_k^{-1}(s))\boldsymbol{z}_k$$

$$+\sum_{k=1}^{L} \boldsymbol{z}_k^T(\boldsymbol{D}_k^{-1}(s) - h_k^{-1}(1-\gamma_k)\boldsymbol{D}_k^{-1}(s)\boldsymbol{J}_{n_k}\boldsymbol{D}_k^{-1}(s))\boldsymbol{z}_k$$

with respect to $\boldsymbol{b}$. This results in the estimator

$$\tilde{\boldsymbol{b}} = \left[\sum_{k=1}^{L} \left\{ \sum_{i \in str_k, i \in s} d_{ki}^{-1} \boldsymbol{x}_{ki} \boldsymbol{x}_{ki}^T - h_k(1-\gamma_k)\bar{\boldsymbol{x}}_{kw}(s)\bar{\boldsymbol{x}}_{kw}^T(s) \right\} \right]^{-1}$$

$$\times \left[\sum_{k=1}^{L} \left\{ \sum_{i \in str_k, i \in s} d_{ki}^{-1} \boldsymbol{x}_{ki} z_i - h_k(1-\gamma_k)\bar{\boldsymbol{x}}_{kw}(s)\bar{z}_{kw} \right\} \right] \quad (4.177)$$

of $\boldsymbol{b}$. Substitution of $\tilde{\boldsymbol{b}}$ in (4.176) leads to the predictor

$$\bar{\boldsymbol{x}}_k^T(s)\tilde{\boldsymbol{b}} + (1-\gamma_k)(\bar{z}_{kw}(s) - \bar{\boldsymbol{x}}_{kw}^T(s)\tilde{\boldsymbol{b}})$$

for $\bar{y}_k(\bar{s})$. This leads to the predictor $\tilde{\boldsymbol{\mu}}^{EB}$ of $\boldsymbol{\mu}$ where

$$\tilde{\mu}_k^{EB} = (1-f_k)\bar{z}_k + f_k[\bar{\boldsymbol{x}}_k^T(\bar{s})\tilde{\boldsymbol{b}} + (1-\gamma_k)(\bar{z}_{kw}(s) - \bar{\boldsymbol{x}}_{kw}^T(s)\tilde{\boldsymbol{b}})]. \quad (4.178)$$

We shall next show that the predictor $\tilde{\boldsymbol{\mu}}^{EB}$ of $\boldsymbol{\mu}$ is also its best unbiased predictor (BUP) or the best linear unbiased predictor (BLUP) under the model given in (4.166). To prove the BUP property, first note that under (4.166), in view of the normality assumption,

$$E[(\tilde{\boldsymbol{\mu}}^{EB} - \boldsymbol{\mu})(\tilde{\boldsymbol{\mu}}^{EB} - \boldsymbol{\mu})^T]$$

$$
\begin{aligned}
&= E[(\tilde{\boldsymbol{\mu}}^{EB} - E(\boldsymbol{\mu}|\boldsymbol{z}))(\tilde{\boldsymbol{\mu}}^{EB} - E(\boldsymbol{\mu}|\boldsymbol{z}))^T] \\
&+ E[(E(\boldsymbol{\mu}|\boldsymbol{z}) - \boldsymbol{\mu})(E(\boldsymbol{\mu}|\boldsymbol{z}) - \boldsymbol{\mu})^T]. \qquad (4.179)
\end{aligned}
$$

Hence, it suffices to show that $\tilde{\boldsymbol{\mu}}^{EB}$ is the BUP of $E(\boldsymbol{\mu}|\boldsymbol{z})$. Arguing as in the previous section, it suffices to show that $\tilde{\mu}_k^{EB}$ is the BUP of $E[\mu_k|\boldsymbol{z}]$. First observe that

$$
\begin{aligned}
E[\mu_k|\boldsymbol{z}] &= (1 - f_k)\bar{z}_k + f_k E(\bar{y}_k(\bar{s})|\boldsymbol{z}) \\
&= (1 - f_k)\bar{z}_k + f_k \bar{\boldsymbol{x}}_k^T(\bar{s})\tilde{\boldsymbol{b}} \\
&+ f_k(1 - \gamma_k)(\bar{z}_{kw}(s) - \bar{\boldsymbol{x}}_{kw}^T(s)\tilde{\boldsymbol{b}}). \qquad (4.180)
\end{aligned}
$$

From definition, it follows that $\tilde{\boldsymbol{b}}$ is an unbiased estimator of $\boldsymbol{b}$. Hence $\tilde{\mu}_k^{EB}$ is an unbiased predictor of $E(\mu_k|\boldsymbol{z})$. To show that this is the BUP, we shall use the result that $\tilde{\mu}_k^{EB}$ is the BUP of $E(\mu_k|\boldsymbol{z})$ if and only if $\tilde{\mu}_k^{EB} - E(\mu_k|\boldsymbol{z})$ is uncorrelated with every $u(\boldsymbol{z})$ such that $E[u(\boldsymbol{z})] \triangleq 0$ for all $\boldsymbol{b}$, σ_v^2 and σ^2. To show this, first compute

$$
\tilde{\mu}_k^{EB} - E(\mu_k|\boldsymbol{z}) = f_k[\bar{\boldsymbol{x}}_k^T(\bar{s})(\tilde{\boldsymbol{b}} - \boldsymbol{b}) - (1 - \gamma_k)\bar{\boldsymbol{x}}_k^T(s)(\tilde{\boldsymbol{b}} - \boldsymbol{b})].
$$

Hence, it suffices to show that

$$
Cov[u(\boldsymbol{z}), \{\bar{\boldsymbol{x}}_k^T(\bar{s}) - (1 - \gamma_k)\bar{\boldsymbol{x}}_k^T(s)\}(\tilde{\boldsymbol{b}} - \boldsymbol{b})] = 0 \qquad (4.181)
$$

for all $\boldsymbol{b}$, σ_v^2 and σ^2. To see this, first note that $E[h(\boldsymbol{z})] = 0$ can be equivalently written as

$$
\begin{aligned}
\mathbf{0} &= \int \cdots \int u(\boldsymbol{z}) \prod_{k=1}^{L} |\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 \boldsymbol{J}_{n_k}|^{-1/2} (2\pi)^{-n_T/2} \\
&\times \exp\left[-\sum_{k=1}^{L} (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\boldsymbol{b})^T (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 \boldsymbol{J}_{n_k})^{-1} \right. \\
&\times \left. (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\boldsymbol{b})/2 \right] d\boldsymbol{z}. \qquad (4.182)
\end{aligned}
$$

Next use the identity

$$
\begin{aligned}
&\sum_{k=1}^{L} (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\boldsymbol{b})^T (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 \boldsymbol{J}_{n_k})^{-1} (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\boldsymbol{b}) \\
&= \sum_{k=1}^{L} (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\tilde{\boldsymbol{b}})^T (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 \boldsymbol{J}_{n_k})^{-1} (\boldsymbol{z}_k - \boldsymbol{X}_k(s)\tilde{\boldsymbol{b}}) \\
&+ (\tilde{\boldsymbol{b}} - \boldsymbol{b})^T \left[\sum_{k=1}^{L} \boldsymbol{X}_k^T(s) (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 \boldsymbol{J}_{n_k})^{-1} \boldsymbol{X}_k(s) \right] (\tilde{\boldsymbol{b}} - \boldsymbol{b}).
\end{aligned}
$$

(4.183)

Using (4.183), and differentiating both sides of (4.182) with respect to $\boldsymbol{b}$, one gets

$$E[u(\boldsymbol{z})\boldsymbol{X}_k(s)(\tilde{\boldsymbol{b}} - \boldsymbol{b})] = 0. \tag{4.184}$$

Since $E[u(\boldsymbol{z})] = 0$ and $P(\boldsymbol{X}_k(s)(\tilde{\boldsymbol{b}} - \boldsymbol{b}) = \boldsymbol{0}) = 0$, one gets

$$Cov(u(\boldsymbol{z}), (\tilde{\boldsymbol{b}} - \boldsymbol{b})) = \boldsymbol{0}. \tag{4.185}$$

Using (4.185), (4.181) follows immediately. This proves the BUP property of $\tilde{\boldsymbol{\mu}}^{EB}$ as an estimator of $\boldsymbol{\mu}$ under the model (4.166).

Next we dispense with the normality assumption, but restrict ourselves to linear unbiased predictors of the form

$$\delta_k(\boldsymbol{z}) = \sum_{l=1}^{L} \sum_{i \in str_k, i \in s} g_{kli} z_i$$

for μ_k, $k = 1, \ldots, L$. Note that the predictor $\tilde{\mu}_k^{EB}$ given in (4.178) is also an unbiased predictor of μ_k. In order to show that this is the BLUP of μ_k, it suffices to show that

$$Cov(\tilde{\mu}_k^{EB} - \mu_k, u(\boldsymbol{z})) = 0, \tag{4.186}$$

for all $\boldsymbol{b}$, σ_v^2 and σ^2 and every u of the form

$$u(\boldsymbol{z}) = \sum_{l=1}^{L} \sum_{i \in str_k, i \in s} a_{kli} z_i$$

satisfying $E(h(\boldsymbol{z})) = 0$ for all $\boldsymbol{b}$, σ_v^2, and σ^2.

To see this, first note that $E(u(\boldsymbol{z})) = 0$ for all $\boldsymbol{b}$ is equivalent to $\sum_{l=1}^{L} \sum_{i \in str_k, i \in s} a_{kli} \boldsymbol{x}_{li} = \boldsymbol{0}$. Next observe that

$$\begin{aligned} \tilde{\mu}_k^{EB} - \mu_k &= f_k[(1 - \gamma_k)\bar{z}_{kw}(s) - \bar{y}_k(\bar{s}) \\ &+ (\bar{\boldsymbol{x}}_k^T(\bar{s}) - (1 - \gamma_k)\bar{\boldsymbol{x}}_{kw}^T)\tilde{\boldsymbol{b}}]. \end{aligned} \tag{4.187}$$

We shall prove the result by showing that

$$Cov[(1 - \gamma_k)\bar{z}_{kw}(s) - \bar{y}_k(\bar{s}), u(\boldsymbol{z})] = 0; \tag{4.188}$$

$$Cov(\tilde{\boldsymbol{b}}, u(\boldsymbol{z})) = \boldsymbol{0}. \tag{4.189}$$

Algebraic simplifications lead to

$$Cov(\bar{z}_{kw}(s), u(\boldsymbol{z})) = h_k^{-1} \left(\sigma^2 + \sigma_v^2 c_k e_k \sum_{i \in str_k, i \in s} a_{kki} \right); \tag{4.190}$$

$$Cov(\bar{y}_k(\bar{s}), u(\boldsymbol{z})) = \sigma_v^2 c_k \sum_{i \in str_k, i \in s} a_{kki}. \tag{4.191}$$

Combining (4.190) and (4.191) one gets

$$\begin{aligned} &Cov[(1-\gamma_k)\bar{z}_{kw}(s) - \bar{y}_k(\bar{s}), u(\boldsymbol{z})] \\ &\quad = \sigma_v^2 c_k \left(\sum_{i \in str_k, i \in s} a_{kki} - \sum_{i \in str_k, i \in s} a_{kki} \right) = 0. \end{aligned} \tag{4.192}$$

This proves (4.188).

To prove (4.189), using the expression for $\tilde{\boldsymbol{b}}$ as given in (4.177), and of $u(\boldsymbol{z})$ it suffices to show that

$$\begin{aligned} \sum_{l=1}^{L} Cov\Bigg[&\sum_{i \in str_l, i \in s} d_{li}^{-1} \boldsymbol{x}_{li} z_i - h_l(1-\gamma_l)\bar{\boldsymbol{x}}_{lw}(s)\bar{z}_{lw}, \\ &\sum_{i \in str_l, i \in s} a_{kli} z_i \Bigg] = 0. \end{aligned} \tag{4.193}$$

But using the definition of γ_l,

$$\begin{aligned} Cov &\left[\sum_{i \in str_l, i \in s} d_{li}^{-1} \boldsymbol{x}_{li} z_i - h_l(1-\gamma_l)\bar{\boldsymbol{x}}_{lw}(s)\bar{z}_{lw}, \sum_{i \in str_l, i \in s} a_{kli} z_i \right] \\ &= \sigma^2 \sum_{i \in str_l, i \in s} a_{kli} \boldsymbol{x}_{li} + \sigma_v^2 c_l h_l \bar{\boldsymbol{x}}_{lw}(s) \sum_{i \in str_l, i \in s} a_{kli} \\ &\quad - h_l(1-\gamma_l)\bar{\boldsymbol{x}}_{lw}(s) h_l^{-1}(\sigma^2 + \sigma_v^2 c_l h_l) \sum_{i \in str_l, i \in s} a_{kli} \\ &= \boldsymbol{0} + (\sigma_v^2 c_l h_l - \sigma_v^2 c_l h_l)\bar{\boldsymbol{x}}_{lw}(s) \sum_{i \in str_l, i \in s} a_{kli} = \boldsymbol{0}. \end{aligned}$$

This proves (4.189).

When σ_v^2 and σ^2 are unknown, we estimate them by the method of moments. First, to estimate σ^2, define $u_{ki} = z_i - \bar{z}_{kw}$ for $i \in str_k$, and $i \in s$. Then the u_{ki} are normal with means $(\boldsymbol{x}_{ki} - \bar{\boldsymbol{x}}_{kw})^T \boldsymbol{b}$, variances $\sigma^2(d_{ki} - h_k^{-1})$ and covariances $-\sigma^2 h_k^{-1}$. Call this variance–covariance matrix $\sigma^2 \boldsymbol{V}_k$. Write $\boldsymbol{S}_k^T = (\boldsymbol{x}_{k1} - \bar{\boldsymbol{x}}_{kw}, \ldots, \boldsymbol{x}_{kn_k} - \bar{\boldsymbol{x}}_{kw})$. Then the BLUE of $\boldsymbol{b}$ is obtained by minimizing $\sum_{k=1}^{L} (\boldsymbol{u}_k - \boldsymbol{S}_k \boldsymbol{b})^T \boldsymbol{V}_k^{-1} (\boldsymbol{u}_k - \boldsymbol{S}_k \boldsymbol{b})$ with respect to $\boldsymbol{b}$, where $\boldsymbol{u}_k$ denotes the vector with elements $z_i - \bar{z}_{kw}$ $(i \in str_k, i \in s)$. The BLUE of $\boldsymbol{b}$ is given

by

$$\hat{\boldsymbol{b}}^* = \left(\sum_{k=1}^{L} \boldsymbol{S}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{S}_k\right)^{-1} \left(\sum_{k=1}^{L} \boldsymbol{S}_k^T \boldsymbol{V}_k^{-1} \boldsymbol{u}_k\right). \tag{4.194}$$

Also, after much simplification,

$$E\left[\sum_{k=1}^{L} (\boldsymbol{u}_k - \boldsymbol{S}_k \hat{\boldsymbol{b}}^*)^T \boldsymbol{V}_k^{-1} (\boldsymbol{u}_k - \boldsymbol{S}_k \hat{\boldsymbol{b}}^*)\right] = \sigma^2 (L - p). \tag{4.195}$$

Hence, the method of moments estimator of σ^2 is given by

$$\hat{\sigma}^2 = (L-p)^{-1} \sum_{k=1}^{L} (\boldsymbol{u}_k - \boldsymbol{S}_k \hat{\boldsymbol{b}}^*)^T \boldsymbol{V}_k^{-1} (\boldsymbol{u}_k - \boldsymbol{S}_k \hat{\boldsymbol{b}}^*). \tag{4.196}$$

Next calculate

$$\begin{aligned} E&\left[\sum_{k=1}^{L} (\boldsymbol{z}_k - \boldsymbol{X}_k \tilde{\boldsymbol{b}})^T (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} (\boldsymbol{z}_k - \boldsymbol{X}_k \tilde{\boldsymbol{b}})\right] \\ &= \sum_{k=1}^{L} tr[(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} E(\boldsymbol{z}_k - \boldsymbol{X}_k \tilde{\boldsymbol{b}})(\boldsymbol{z}_k - \boldsymbol{X}_k \tilde{\boldsymbol{b}})^T] \\ &= \sum_{k=1}^{L} tr[(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} V(\boldsymbol{z}_k - \boldsymbol{X}_k \tilde{\boldsymbol{b}})]. \end{aligned} \tag{4.197}$$

Write

$$V(\boldsymbol{z}_k - \boldsymbol{X}_k \tilde{\boldsymbol{b}}) = V(\boldsymbol{z}_k) + V(\boldsymbol{X}_k \tilde{\boldsymbol{b}}) - Cov(\boldsymbol{z}_k, \boldsymbol{X}_k \tilde{\boldsymbol{b}}) - Cov(\boldsymbol{X}_k \tilde{\boldsymbol{b}}, \boldsymbol{z}_k). \tag{4.198}$$

Now

$$V(\boldsymbol{z}_k) = \sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k}; \tag{4.199}$$

$$V(\tilde{\boldsymbol{b}}) = \left(\sum_{k=1}^{L} \boldsymbol{X}_k^T (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} \boldsymbol{X}_k\right)^{-1}. \tag{4.200}$$

Hence,

$$\begin{aligned} &\sum_{k=1}^{L} tr[(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} V(\boldsymbol{X}_k \tilde{\boldsymbol{b}})] \\ &\quad = tr\left[\left\{\sum_{k=1}^{L} \boldsymbol{X}_k^T (\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} \boldsymbol{X}_k)\right\}^{-1}\right. \end{aligned}$$

$$\times \quad \sum_{k=1}^{L} \boldsymbol{X}_k^T(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} \boldsymbol{X}_k \Big]$$

$$= \quad tr(\boldsymbol{I}_p) = p. \tag{4.201}$$

Finally,

$$\begin{aligned} Cov(\tilde{\boldsymbol{b}}, \boldsymbol{z}_k) &= \left(\sum_{k=1}^{L} \boldsymbol{X}_k^T(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} \boldsymbol{X}_k \right)^{-1} \\ &\times \quad Cov(\boldsymbol{X}_k^T(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} \boldsymbol{z}_k, \boldsymbol{z}_k) \\ &= \left(\sum_{k=1}^{L} \boldsymbol{X}_k^T(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} \boldsymbol{X}_k \right)^{-1} \boldsymbol{X}_k^T. \end{aligned}$$

Thus, after much algebraic simplification,

$$\sum_{k=1}^{L} tr[(\sigma^2 \boldsymbol{D}_k(s) + \sigma_v^2 c_k \boldsymbol{J}_{n_k})^{-1} Cov(\boldsymbol{X}_k \tilde{\boldsymbol{b}}, \boldsymbol{z}_k)] = p. \tag{4.202}$$

From (4.197)–(4.202), using the method of moments, one gets

$$\sum_{k=1}^{L} (\boldsymbol{z}_k - \boldsymbol{X}_k \tilde{\boldsymbol{b}})^T (\hat{\sigma}^2 \boldsymbol{D}_k(s) + \hat{\sigma}_v^2 c_k \boldsymbol{J}_{n_k})^{-1} (\boldsymbol{z}_k - \boldsymbol{X}_k \tilde{\boldsymbol{b}}) = L - p. \tag{4.203}$$

Using (4.177), (4.196) and (4.203), and solving iteratively, one gets the estimators $\hat{\boldsymbol{b}}$, $\hat{\sigma}_v^2$ and $\hat{\sigma}^2$ of $\boldsymbol{b}$, σ_v^2 and σ^2 respectively. Substitution of these estimators in (4.180) where $\hat{\boldsymbol{b}}$ replaces $\tilde{\boldsymbol{b}}$ leads to the final EB estimator $\hat{\mu}_k^{EB}$ of μ_k, for all $k = 1, \ldots, l$. The resulting EB estimator $\hat{\boldsymbol{\mu}}^{EB}$ estimates $\boldsymbol{\mu}$.

CHAPTER 5

Hierarchical Bayes estimation

5.1 Introduction

In the previous chapter, we considered empirical Bayes (EB) estimation of stratum means with or without the presence of auxiliary information. The results were given both under the assumption of normality, and under the assumption of posterior linearity, but without any further distributional assumptions. The present chapter will consider an alternate analysis based on hierarchical Bayes (HB) models. We repeat our earlier comment that the two approaches lead to comparable point estimates. However, the HB method has the advantage of reporting simple measures of standard errors associated with these estimates (usually the posterior means), namely the posterior standard deviations.

The outline of the remaining sections is as follows. In Section 5.2 we show an HB approach often achieves a synthesis between model- and design-based estimators, and how some of the EB estimators derived in Section 4.2 can also be viewed as HB estimators. In Section 5.3, we consider HB estimates of stratum means without the use of auxiliary information. The results of this section will, therefore, be comparable to those of Section 4.3 or 4.4. We shall provide conditions under which the two procedures provide identical point estimates, but point out at the same time why a naive EB procedure will typically underestimate the actual standard error. In Section 5.4, we shall develop HB procedures analogous to the EB procedure of Section 4.6. In addition to the general theory, special cases will be considered where estimators obtained by the two methods are identical, and where these estimators also turn out to be the best linear unbiased predictors, and best unbiased predictors under the added normality assumption. In Section 5.5, we shall consider HB estimation under general mixed linear models. Finally, in Section 5.6, we shall consider HB methods for simultaneous estimation of stratum means based on generalized linear models.

5.2 Stepwise Bayes estimators

In the previous chapter, we showed how several stepwise Bayes estimators can also be viewed as empirical Bayes estimators. In this chapter, we provide a hierarchical Bayes derivation of the same. In addition, we want to highlight a possible synthesis between model- and design-based estimators.

In finite population sampling, often a distinction is made between model- and design-based estimators of the parameters of interest. The model-based estimators depend on the estimated parameters of the model, while the design-based estimators depend on the known selection probabilities of the different units of the population. Little (1983) mentions that although the two approaches lead to similar point and interval esimators for the population mean under simple random sampling and a normal model, the two approaches can yield markedly different results especially for unequal probability sampling. We shall see in this section that often it is possible to achieve a synthesis between the two approaches in the context of point estimation, although the corresponding standard errors and interval estimators can substantially differ.

First, a HB approach is taken to generate estimators of the population total. The calculations are taken from Ghosh and Sinha (1990). The approach is related to the EB approach of the previous section, and is also related to the prediction approach of Royall (1970). However, the HB approach has one advantage over the other two methods in that unlike the EB or the prediction approach, it provides natural measures of accuracy, namely the posterior standard errors associated with the parameters of interest.

In order to introduce the HB method, we begin with a finite population with units labelled $1, \ldots, N$. Also, as before, s and $\bar{s}$ denote the sampled and the unsampled units respectively. We denote by $y_1, \ldots, y_N$ the characteristics of interest associated with the N units in the population. Consider the Bayes model according to which conditional on θ, the y_i are independent $N(\theta a_i, \sigma_i^2)$, where the a_i and the $\sigma_i^2 > 0$ are known constants. It is also assumed that marginally θ is uniform over the the real line. The following theorem provides the joint posterior of the unobserved y_i given the observed sample z. For notational simplicity, it is assumed that units $1, \ldots, n$ are observed in the sample.

Theorem 5.1 *Given* $z = (z_1, \ldots, z_n)^T$, *the joint posterior distri-*

bution of $y_{n+1}, \ldots, y_N$ is multivariate normal with

$$E(y_j|\boldsymbol{z}) = \left(\sum_{i=1}^{n} a_i\sigma_i^{-2}z_i / \sum_{i=1}^{n} a_i\sigma_i^{-2}\right) a_j; \tag{5.1}$$

$$V(y_j|\boldsymbol{z}) = \sigma_j^2 + a_j^2 / \sum_{i=1}^{n} a_i^2\sigma_i^{-2}; \tag{5.2}$$

$$Cov(y_j, y_k|\boldsymbol{z}) = a_j a_k / \sum_{i=1}^{n} a_i^2\sigma_i^{-2}, \tag{5.3}$$

where $n+1 \leq j \neq k \leq N$.

Proof. The joint (improper) pdf of $y_1, \ldots, y_N$ and θ is

$$f(y_1, \ldots, y_N, \theta) \propto \exp\left[-\sum_{i=1}^{N}(y_i - \theta a_i)^2/2\right]. \tag{5.4}$$

Writing $\boldsymbol{y}(\bar{s}) = (y_{n+1}, \ldots, y_N)^T$, the above pdf can be written as

$$f(\boldsymbol{z}, \boldsymbol{y}(\bar{s})) \propto exp[-\{\boldsymbol{y}(\bar{s})^T \boldsymbol{A}_{22}\boldsymbol{y}(\bar{s}) + \boldsymbol{z}^T\boldsymbol{A}_{11}\boldsymbol{z} - 2\boldsymbol{y}(\bar{s})^T\boldsymbol{A}_{21}\boldsymbol{z}\}], \tag{5.5}$$

where

$$\begin{aligned} \boldsymbol{A}_{11} &= Diag(\sigma_1^{-2}, \ldots, \sigma_n^{-2}) \\ &- (a_1\sigma_1^{-2}, \ldots, a_n\sigma_n^{-2})^T(a_1\sigma_1^{-2}, \ldots, a_n\sigma_n^{-2}) / \sum_{i=1}^{N} a_i^2\sigma_i^{-2} \end{aligned} \tag{5.6}$$

$$\begin{aligned} \boldsymbol{A}_{22} &= Diag(\sigma_{n+1}^{-2}, \ldots, \sigma_N^{-2}) - \left(\sum_{i=1}^{N} a_i^2\sigma_i^{-2}\right)^{-1} \\ &- (a_{n+1}\sigma_{n+1}^{-2}, \ldots, a_N\sigma_N^{-2})^T(a_{n+1}\sigma_{n+1}^{-2}, \ldots, a_N\sigma_N^{-2}) \end{aligned} \tag{5.7}$$

and

$$\boldsymbol{A}_{21} = (a_{n+1}\sigma_{n+1}^{-2}, \ldots, a_N\sigma_N^{-2})^T(a_1\sigma_1^{-2}, \ldots, a_n\sigma_n^{-2}) / \sum_{i=1}^{N} a_i^2\sigma_i^{-2}. \tag{5.8}$$

Hence, the joint posterior of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$ is $N(\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{z}, \boldsymbol{A}_{22}^{-1})$. Using the familiar matrix inversion formula

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}/(1 + \boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}),$$

one gets after some simplifications

$$A_{22}^{-1} = Diag(\sigma_1^2, \ldots, \sigma_n^2) + (a_{n+1}, \ldots, a_N)^T (a_{n+1}, \ldots, a_N) / \sum_{i=1}^{n} a_i^2 \sigma_i^{-2} \quad (5.9)$$

and

$$A_{22}^{-1} A_{21} z = \left\{ \sum_{i=1}^{n} a_i \sigma_i^{-2} z_i / a_i^2 \sigma_i^{-2} \right\} (a_{n+1}, \ldots, a_N)^T. \quad (5.10)$$

The proof of the theorem is complete. □

Based on the above theorem, the posterior mean of

$$\mu = N^{-1} \sum_{i=1}^{N} y_i$$

is given by

$$E(\mu|z) = N^{-1} \left[n\bar{z} + \left(\sum_{i=1}^{n} a_i \sigma_i^{-2} z_i / \sum_{i=1}^{n} a_i^2 \sigma_i^{-2} \right) \sum_{j=n+1}^{N} a_j \right], \quad (5.11)$$

where $\bar{z} = n^{-1} \sum_{i=1}^{n} z_i$. We find that this estimator is identical to the EB estimator derived in Section 4.2. As special cases, we derived in Section 4.2 the Horvitz–Thompson estimator, as well as several other estimators of Royall and Basu.

Since θ has an improper distribution, the unconditional Bayes risk of every estimator of the population mean is infinite. However, conditional on θ, the Bayes risk of the estimator of the population mean derived in (5.11) is given by

$$N^{-2} E \left[\left(\sum_{i \in s} a_i \sigma_i^{-2} z_i / \sum_{i \in s} a_i^2 \sigma_i^{-2} \right) \sum_{j \in \bar{s}} a_j - \sum_{j \in \bar{s}} y_j \right]^2$$

$$= \sum_{s \in S} p(s) \left[\left(\sum_{j \in \bar{s}} a_j \right)^2 / \left(\sum_{i \in s} a_i^2 \sigma_i^{-2} \right) + \sum_{j \in \bar{s}} \sigma_j^2 \right], \quad (5.12)$$

where $p(s)$ denotes the probability of selecting sample s, and S denotes the set of all possible samples. Thus, using a Bayesian approach, one selects those units i for which the above conditional Bayes risk is minimized.

In the special case, when $a_i = \pi_i$ and $\sigma_i^2 = \pi_i/(1-\pi_i)$, the Bayes estimator of the finite population mean is the Horvitz–Thompson estimator. The corresponding conditional Bayes risk is

$$\sum_{s \in S} p(s) \sum_{j \in \bar{s}} \pi_j/(1-\pi_j).$$

It follows then that since $\pi_j/(1-\pi_j)$ is strictly increasing in π_j, one should select those units with the largest π_j. This seems very sensible because units with larger coefficients of variation should have bigger probabilities of being selected in the sample, inasmuch as these are hard to predict from the values of the remaining units.

We now examine the implication of the above result in the context of the famous circus example of Basu (1971). First, we reproduce Basu's example.

'The circus owner is planning to ship his 50 adult elephants and so he needs a rough estimate of the total weight of elephants. As weighing an elephant is a cumbersome process, the owner wants to estimate the total weight by weighing just one elephant. Which elephant should he weigh? So the owner looks back on his records and discovers a list of the elephants' weights taken 3 years ago. He finds that Sambo, the middle-sized elephant was the average (in weight) elephant in his herd. He checks with the elephant trainer who reassures him (the owner) that Sambo may still be considered to be the average elephant in the herd. Therefore, the owner plans to weigh Sambo and take $50y$ (where y is the present weight of Sambo) as an estimate of the total weight $Y = Y_1 + \cdots + Y_{50}$ of the 50 elephants. But the circus statistician is horrified when he learns of the owner's purposive sampling plan. "How can you get an unbiased estimate of Y this way?" protests the statistician. So, together they work out a compromise sampling plan. With the help of a table of random numbers they devise a plan that allots a selection probability of 99/100 to Sambo and equal selection probabilities of 1/4900 to each of the other 49 elephants. Naturally, Sambo is selected and the owner is happy. "How are you going to estimate Y?" asks the statistician. "Why? The estimate ought to be $50y$ of course," says the owner. "Oh! No! That cannot possibly be right," says the statistician. "I recently read an article in the Annals of Mathematical Statistics where it is proved that the Horvitz–Thompson estimator is the unique hyperadmissible estimator in the class of all generalized polynomial unbiased estimators." "What is the Horvitz–Thompson estimate in this case?"

asks the owner, duly impressed. "Since the selection probability for Sambo in our plan was 99/100," says the statistician, "the proper estimate of Y is $100y/99$ and not $50y$". "And, how would you have estimated Y?" inquires the incredulous owner, "if our sampling plan made us select, say, the big elephant Jumbo?" "According to what I understand of the Horvitz–Thompson estimation method," says the unhappy statistician, "the proper estimate of Y would then have been $4900y$, where y is Jumbo's weight." That is how the statistician lost his circus job (and perhaps became a teacher of statistics!).'

Basu's example clearly demonstrates that unless one uses the selection probabilities judiciously, the Horvitz–Thompson estimate can be meaningless. Our result suggests that if the circus statistician has to select only one elephant, he should select Jumbo, the big-sized elephant with probability 1. Even if one does not accept this extreme view of purposive sampling, assigning probability 1/4900 to Jumbo is clearly wrong, and it is no wonder that the circus statistician lost his job. Also, if one accepts our view on the selection probabilities of the different units, the Horvitz–Thompson estimator also becomes sensible, because the weights are inversely proportional to the selection probabilities.

One of the advantages of the HB approach is that the posterior variances can also be easily calculated from Theorem 5.1. Thus, it is easy to construct the credible intervals using the normal percentiles. The EB approach typically underestimates the variability, whereas the least squares prediction, in the absence of any distributional assumptions, also does not yield any confidence intervals.

5.3 Estimation of stratum means

We shall continue with the notations of the previous chapter. Consider the Bayesian model given in (I) and (II) on page 165 of Section 4.3. However, in this case, rather than estimating m, σ^2 and τ^2 from the data, we assign some priors (often diffuse) to these parameters.

First consider the case when $\lambda = \sigma^2/\tau^2$ is known, but one assigns a third stage to the hierarchical model under which

(III) m and r are independent with $m \sim$ uniform $(-\infty, \infty)$, while $R = \sigma^{-2} \sim \text{gamma}(a/2, b/2)$.

Specific choices of a and b will be given later in this section. We

allow the possibility of $a = 0$, $b = 0$ or both as long as the posterior distribution of $y(\bar{s})$ given s and $\boldsymbol{z}$ remains proper.

The first theorem of this section gives the predictive distribution of $\boldsymbol{y}(\bar{s})$ given s and $\boldsymbol{z}$ when λ is known. We shall see that some of the Bayesian inference procedures in this case agree perfectly with some classical procedures as well as with some of the EB methods developed in the previous chapter. Recall the notations $\bar{z}_k = n_k^{-1} \sum_{i \in str_k} z_i$, $k = 1, \ldots, L$, $\bar{z} = \sum_{k=1}^{L} n_k \bar{z}_k / n_T$, $n_T = \sum_{k=1}^{L} n_k$. We shall denote the Kronecker sum by $\bigoplus$, i.e. $\bigoplus_{k=1}^{L} \boldsymbol{A}_k$ = block diagonal $(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L)$.

Theorem 5.2 *Consider the hierarchical model given in (I) and (II) of Section 4.3 and (III) of the present section. Then the predictive distribution of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$ is multivariate-t with location vector*

$$[(1 - B_1)\bar{z}_1 + B_1 \bar{z}, \ldots, (1 - B_L)\bar{z}_L + B_L \bar{z}]^T,$$

scale matrix $(n_T + b - 1)^{-1}(a + Q(\boldsymbol{z}))\boldsymbol{G}$, and degrees of freedom $n_T + b - 1$, where

$$Q(\boldsymbol{z}) = \sum_{k=1}^{L} \sum_{i \in str_k} (y_i - \bar{z}_k)^2 + \lambda \sum_{k=1}^{L} (1 - B_k)(\bar{z}_k - \bar{z})^2,$$

and

$$\begin{aligned} \boldsymbol{G} &= \bigoplus_{k=1}^{L} [\boldsymbol{I}_{N_k - n_k} + (\lambda + n_k)^{-1} \boldsymbol{J}_{N_k - n_k}] \\ &+ \left[\sum_{k=1}^{L} (1 - B_k) \right]^{-1} [B_1 \mathbf{1}_{N_1 - n_1}^T, \ldots, B_L \mathbf{1}_{N_L - n_L}^T]^T \\ &\times [B_1 \mathbf{1}_{N_1 - n_1}^T, \ldots, B_L \mathbf{1}_{N_L - n_L}^T]. \end{aligned}$$

Proof. First write the joint pdf of $\boldsymbol{z}, \boldsymbol{y}(\bar{s}), \boldsymbol{\theta}, m$ and r as

$$\begin{aligned} &f(\boldsymbol{z}, \boldsymbol{y}(\bar{s}), \boldsymbol{\theta}, m, r) \\ &\quad \propto r^{N_T/2} \exp\left[-(r/2) \left\{ \sum_{k=1}^{L} \sum_{i \in str_k} (y_i - \theta_k)^2 \right.\right. \\ &\quad \left.\left. + \sum_{k=1}^{L} \sum_{j \in str_k, j \in \bar{s}} (y_j - \theta_k)^2 \right\} \right] (\lambda r)^{L/2} \end{aligned}$$

$$\times \exp\left[-\lambda r \sum_{k=1}^{L}(\theta_k - m)^2/2\right] \exp(-ar/2) r^{b/2-1}. \quad (5.13)$$

Integrate first with respect to m. Then the joint (improper) pdf of $\boldsymbol{z}$, $\boldsymbol{y}(\bar{s})$, $\boldsymbol{\theta}$ and r is given by

$$\begin{aligned} & f(\boldsymbol{z}, \boldsymbol{y}(\bar{s}), \boldsymbol{\theta}, r) \\ & \quad \propto r^{(N_T-1)/2} \exp\left[-r\left\{\sum_{k=1}^{L}\sum_{i\in str_k}(y_i-\theta_k)^2 \right.\right. \\ & \quad \left.\left. + \sum_{k=1}^{L}\sum_{j\in str_k, j\in\bar{s}}(y_j-\theta_k)^2\right\}/2\right](\lambda r)^{(L-1)/2} \\ & \quad \times \exp\left[-\lambda r \sum_{k=1}^{L}(\theta_k-\bar{\theta})^2\right] \exp(-ar/2) r^{b/2-1}. \quad (5.14) \end{aligned}$$

Next write

$$\begin{aligned} & \sum_{k=1}^{L}\sum_{i\in str_k}(y_i-\theta_k)^2 + \sum_{k=1}^{L}\sum_{j\in str_k, j\in\bar{s}}(y_j-\theta_k)^2 + \lambda\sum_{k=1}^{L}(\theta_k-\bar{\theta})^2 \\ & \quad = \boldsymbol{\theta}^T[\boldsymbol{D} + \lambda(\boldsymbol{I}_L - L^{-1}\boldsymbol{J}_L)]\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\left[\bigoplus_{k=1}^{L}\mathbf{1}_{n_k}^T\boldsymbol{z} + \bigoplus_{k=1}^{L}\mathbf{1}_{n_k}^T\boldsymbol{y}(\bar{s})\right] \\ & \quad + \boldsymbol{z}^T\boldsymbol{z} + \boldsymbol{y}^T(\bar{s})\boldsymbol{y}(\bar{s}), \quad (5.15) \end{aligned}$$

where $\boldsymbol{D} = Diag(N_1, \ldots, N_L)$, $\boldsymbol{I}_L$ is the identity matrix of order L, $\mathbf{1}_u$ is a u-component column-vector with element equal to 1, $\boldsymbol{J}_u = \mathbf{1}_u\mathbf{1}_u^T$. Write $\boldsymbol{E} = \boldsymbol{D} + \lambda(\boldsymbol{I}_L - L^{-1}\boldsymbol{J}_L)$, and $\boldsymbol{w} \equiv \boldsymbol{w}(\boldsymbol{z}, \boldsymbol{y}(\bar{s})) = (\bigoplus_{k=1}^{L}\mathbf{1}_{n_k}^T)\boldsymbol{z} + (\bigoplus_{k=1}^{L}\mathbf{1}_{N_k-n_k}^T)\boldsymbol{y}(\bar{s})$. Then the right-hand side of (5.15) simplifies to

$$(\boldsymbol{\theta} - \boldsymbol{E}^{-1}\boldsymbol{w})^T\boldsymbol{E}(\boldsymbol{\theta} - \boldsymbol{E}^{-1}\boldsymbol{w}) + \boldsymbol{z}^T\boldsymbol{z} + \boldsymbol{y}^T(\bar{s})\boldsymbol{y}(\bar{s}) - \boldsymbol{w}^T\boldsymbol{E}^{-1}\boldsymbol{w}. \quad (5.16)$$

Now integrating with respect to $\boldsymbol{\theta}$, it follows from (5.16) that the joint (improper) pdf of $\boldsymbol{z}$, $\boldsymbol{y}(\bar{s})$ and r is given by

$$\begin{aligned} f(\boldsymbol{z}, \boldsymbol{y}(\bar{s}), r) \quad & \propto \quad r^{(N_T+b-1)/2-1}|\boldsymbol{E}|^{-1/2}\lambda^{(L-1)/2} \\ & \times \quad exp[-r(a + \boldsymbol{z}^T\boldsymbol{z} + \boldsymbol{y}^T(\bar{s})\boldsymbol{y}(\bar{s}) - \boldsymbol{w}^T\boldsymbol{E}^{-1}\boldsymbol{w})/2]. \end{aligned} \quad (5.17)$$

Now integrating with respect to r, it follows from (5.17) that the joint (improper) pdf of $\boldsymbol{z}$ and $\boldsymbol{y}(\bar{s})$ is given by

$$f(\boldsymbol{z}, \boldsymbol{y}(\bar{s})) \quad \propto \quad |\boldsymbol{E}|^{-1/2} M^{(L-1)/2}$$

$$\times \quad [a + \boldsymbol{z}^T\boldsymbol{z} + \boldsymbol{y}^T(\bar{s})\boldsymbol{y}(\bar{s}) - \boldsymbol{w}^T\boldsymbol{E}^{-1}\boldsymbol{w}]^{-(N_T+b-1)/2}. \tag{5.18}$$

Next simplify

$$\begin{aligned} &\boldsymbol{z}^T\boldsymbol{z} + \boldsymbol{y}^T(\bar{s})\boldsymbol{y}(\bar{s}) - \boldsymbol{w}^T\boldsymbol{E}^{-1}\boldsymbol{w} \\ &= \boldsymbol{z}^T\boldsymbol{F}_{11}\boldsymbol{z} + \boldsymbol{y}^T(\bar{s})\boldsymbol{F}_{22}\boldsymbol{y}(\bar{s}) - 2\boldsymbol{y}^T(\bar{s})\boldsymbol{F}_{21}\boldsymbol{z}, \end{aligned} \tag{5.19}$$

where

$$\boldsymbol{F}_{11} = \boldsymbol{I}_{n_T} - \left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k}\right) \boldsymbol{E}^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k}^T\right); \tag{5.20}$$

$$\boldsymbol{F}_{22} = \boldsymbol{I}_{N_T-n_T} - \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}\right) \boldsymbol{E}^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}^T\right); \tag{5.21}$$

$$\boldsymbol{F}_{21} = \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}\right) \boldsymbol{E}^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k}^T\right). \tag{5.22}$$

Also, one can write the right-hand side of (5.19) as

$$\begin{aligned} &[\boldsymbol{y}(\bar{s}) - \boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}\boldsymbol{z}]^T\boldsymbol{F}_{22} \\ &\quad \times [\boldsymbol{y}(\bar{s}) - \boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}\boldsymbol{z}] + \boldsymbol{z}^T(\boldsymbol{F}_{11} - \boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21})\boldsymbol{z}. \end{aligned} \tag{5.23}$$

From (5.18), (5.19) and (5.23), it follows that the predictive density of $\boldsymbol{y}(\bar{s})$ given z is multivariate-t with location-vector $\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}\boldsymbol{z}$, scale-matrix $(n_T + b - 1)^{-1}[a + \boldsymbol{z}^T(\boldsymbol{F}_{11} - \boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21})\boldsymbol{z}]\boldsymbol{F}_{22}^{-1}$ and degrees of freedom $n_T + b - 1$.

It remains to find $\boldsymbol{F}_{22}^{-1}$, $\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}$ and $\boldsymbol{F}_{11} - \boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}$. First, using Exercise 2.9, p. 33, of Rao (1973), one gets

$$\begin{aligned} \boldsymbol{F}_{22}^{-1} &= \left[\boldsymbol{I}_{N_T-n_T} - \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}\right) \boldsymbol{E}^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}^T\right)\right]^{-1} \\ &= \boldsymbol{I}_{N_T-n_T} + \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}\right) \\ &\quad \times \left(\boldsymbol{E} - \bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}^T \mathbf{1}_{N_k-n_k}\right)^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}^T\right) \\ &= \boldsymbol{I}_{N_T-n_T} + \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}\right) \end{aligned}$$

$$\times [Diag(\lambda + n_1, \ldots, \lambda + n_L) - \lambda L^{-1} \boldsymbol{J}_L]^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}^T_{N_k - n_k} \right). \tag{5.24}$$

Next, using Exercise 2.8, p. 33, of Rao (1973), one gets

$$\begin{aligned} &[Diag(\lambda + n_1, \ldots, \lambda + n_L) - \lambda L^{-1} \boldsymbol{J}_L]^{-1} \\ &\quad = \lambda^{-1} Diag(B_1, \ldots, B_L) \\ &\quad + \left(\sum_{k=1}^{L} (1 - B_k) \right)^{-1} \lambda^{-1} (B_1, \ldots, B_L)^T (B_1, \ldots, B_L). \end{aligned} \tag{5.25}$$

Combining (5.24) and (5.25), one gets

$$\begin{aligned} \boldsymbol{F}_{22}^{-1} &= \boldsymbol{I}_{N_T - n_T} + \bigoplus_{k=1}^{L} (\lambda + n_k)^{-1} \boldsymbol{J}_{N_k - n_k} \\ &\quad + \left(\sum_{k=1}^{L} (1 - B_k) \right)^{-1} [B_1 \mathbf{1}^T_{N_1 - n_1}, \ldots, B_L \mathbf{1}^T_{N_L - n_L}]^T \\ &\quad \times [B_1 \mathbf{1}^T_{N_1 - n_1}, \ldots, B_L \mathbf{1}^T_{N_L - n_L}] \\ &= \bigoplus_{k=1}^{L} [\boldsymbol{I}_{N_k - n_k} + (\lambda + n_k)^{-1} \boldsymbol{J}_{N_k - n_k}] \\ &\quad + \lambda^{-1} \left(\sum_{k=1}^{L} (1 - B_k) \right)^{-1} [B_1 \mathbf{1}^T_{N_1 - n_1}, \ldots, B_L \mathbf{1}^T_{N_L - n_l}]^T \\ &\quad \times [B_1 \mathbf{1}^T_{N_1 - n_1}, \ldots, B_L \mathbf{1}^T_{N_L - n_L}]. \end{aligned} \tag{5.26}$$

Next observe that

$$\begin{aligned} \boldsymbol{E}^{-1} &= Diag((\lambda + n_1)^{-1}, \ldots, (\lambda + n_L)^{-1}) \\ &+ \lambda \left(\sum_{k=1}^{L} N_k (\lambda + N_k)^{-1} \right)^{-1} \\ &\times [(\lambda + N_1)^{-1}, \ldots, (\lambda + N_L)^{-1}]^T \\ &\times [(\lambda + N_1)^{-1}, \ldots, (\lambda + N_L)^{-1}], \end{aligned} \tag{5.27}$$

so that

$$\begin{aligned}
\boldsymbol{F}_{21} &= \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}\right) \boldsymbol{E}^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k}^T\right) \\
&= \bigoplus_{k=1}^{L} (\lambda + N_k)^{-1} \mathbf{1}_{N_k-n_k} \mathbf{1}_{n_k}^T + \lambda \left[\sum_{k=1}^{L} N_k(\lambda + N_k)^{-1}\right]^{-1} \\
&\times [(\lambda + N_1)^{-1} \mathbf{1}_{N_1-n_1}^T, \ldots, (\lambda + N_L)^{-1} \mathbf{1}_{N_L-n_L}^T] \\
&\times [(\lambda + N_1)^{-1} \mathbf{1}_{n_1}^T, \ldots, (\lambda + N_L)^{-1} \mathbf{1}_{n_L}^T]. \qquad (5.28)
\end{aligned}$$

Combining (5.26) and (5.28), after much algebraic simplification one gets,

$$\begin{aligned}
\boldsymbol{F}_{22}^{-1} \boldsymbol{F}_{21} &= \bigoplus_{k=1}^{L} (\lambda + n_k)^{-1} \mathbf{1}_{N_k-n_k} \mathbf{1}_{n_k}^T \\
&+ \left(\sum_{k=1}^{L} (1 - B_k)\right)^{-1} [B_1 \mathbf{1}_{N_1-n_1}^T, \ldots, B_L \mathbf{1}_{N_L-n_L}^T]^T \\
&\times [(\lambda + n_1)^{-1} \mathbf{1}_{n_1}^T, \ldots, (\lambda + n_L)^{-1} \mathbf{1}_{n_L}^T]. \qquad (5.29)
\end{aligned}$$

Hence, from (5.29),

$$\boldsymbol{F}_{22}^{-1} \boldsymbol{F}_{21} \boldsymbol{z} = [(1 - B_1)\bar{z}_1 + B_1 \bar{z}, \ldots, (1 - B_L)\bar{z}_L + B_L \bar{z}]^T. \quad (5.30)$$

Next find

$$\begin{aligned}
\boldsymbol{F}_{12} \boldsymbol{F}_{22}^{-1} \boldsymbol{F}_{21} = &\left[\left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k}\right) \boldsymbol{E}^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k-n_k}^T\right)\right] \\
&\times \left[\bigoplus_{k=1}^{L} (\lambda + n_k)^{-1} \mathbf{1}_{N_k-n_k} \mathbf{1}_{n_k}^T + \left(\sum_{k=1}^{L} (1 - B_k)\right)^{-1}\right. \\
&\times [B_1 \mathbf{1}_{N_1-n_1}^T, \ldots, B_L \mathbf{1}_{N_L-n_L}^T]^T \\
&\left.\times [(\lambda + n_1)^{-1} \mathbf{1}_{n_1}^T, \ldots, (\lambda + n_L)^{-1} \mathbf{1}_{n_L}^T]\right] \\
= &\left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k}\right) \boldsymbol{E}^{-1} \left(\bigoplus_{k=1}^{L} (N_k - n_k)(\lambda + n_k)^{-1} \mathbf{1}_{n_k}^T\right) \\
&+ \left(\sum_{k=1}^{L} (1 - B_k)\right)^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k} \boldsymbol{E}^{-1}\right) \\
&\times [(N_1 - n_1)B_1, \ldots, (N_L - n_L)B_L]^T
\end{aligned}$$

$$\times [(\lambda + n_1)^{-1}\mathbf{1}_{n_1}^T, \ldots, (\lambda + n_L)^{-1}\mathbf{1}_{n_L}^T]. \tag{5.31}$$

Now, using (5.27), it follows after some algebraic simplification that

$$\begin{aligned}
&\left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k}\right) \boldsymbol{E}^{-1} \left(\bigoplus_{k=1}^{L} (N_k - n_k)(\lambda + N_k)^{-1}\mathbf{1}_{n_k}^T\right) \\
&\quad = \bigoplus_{k=1}^{L} \{(\lambda + n_k)^{-1} - (\lambda + N_k)^{-1}\}\boldsymbol{J}_{n_k} \\
&\quad + \lambda \left[\sum_{k=1}^{L} N_k(\lambda + N_k)^{-1}\right]^{-1} \\
&\quad \times [(\lambda + N_1)^{-1}\mathbf{1}_{n_1}^T, \ldots, (\lambda + N_L)^{-1}\mathbf{1}_{n_L}^T]^T \\
&\quad \times [(\lambda + n_1)^{-1}\mathbf{1}_{n_1}^T, \ldots, (\lambda + n_L)^{-1}\mathbf{1}_{n_L}^T] \\
&\quad - \lambda \left(\sum_{k=1}^{L} N_k(\lambda + N_k)^{-1}\right)^{-1} \\
&\quad \times [(\lambda + N_1)^{-1}\mathbf{1}_{n_1}^T, \ldots, (\lambda + N_L)^{-1}\mathbf{1}_{n_L}^T] \\
&\quad \times [(\lambda + N_1)^{-1}\mathbf{1}_{n_1}^T, \ldots, (\lambda + N_L)^{-1}\mathbf{1}_{n_L}^T].
\end{aligned} \tag{5.32}$$

Again, after much simplification,

$$\begin{aligned}
&\left[\sum_{k=1}^{L}(1 - B_k)\right]^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{n_k}\right) \boldsymbol{E}^{-1} \\
&\quad \times [(N_1 - n_1)B_1, \ldots, (N_L - n_L)B_L]^T \\
&\quad \times [(\lambda + n_1)^{-1}\mathbf{1}_{n_1}^T, \ldots, (\lambda + n_L)^{-1}\mathbf{1}_{n_L}^T] \\
&\quad = \lambda \left[\sum_{k=1}^{L}(1 - B_k)\right]^{-1} \\
&\quad \times [\{(\lambda + n_1)^{-1} - (\lambda + N_1)^{-1}\}\mathbf{1}_{n_1}^T, \\
&\qquad \ldots, \{(\lambda + n_L)^{-1} - (\lambda + N_L)^{-1}\}\mathbf{1}_{n_L}^T] \\
&\quad \times [(\lambda + n_1)^{-1}\mathbf{1}_{n_1}^T, \ldots, (\lambda + n_L)^{-1}\mathbf{1}_{n_L}^T] \\
&\quad + \lambda \left[\left\{\sum_{k=1}^{L}(1 - B_k)\right\}^{-1} - \left\{\sum_{k=1}^{L} N_k(\lambda + N_k)^{-1}\right\}^{-1}\right] \\
&\quad \times [(\lambda + N_1)^{-1}\mathbf{1}_{n_1}^T, \ldots, (\lambda + N_L)^{-1}\mathbf{1}_{n_L}^T]^T
\end{aligned}$$

$$\times \quad [(\lambda+n_1)^{-1}\mathbf{1}_{n_1}^T,\ldots,(\lambda+n_L)^{-1}\mathbf{1}_{n_L}^T]. \tag{5.33}$$

Also, observe that from (5.20) and (5.27),

$$\begin{aligned}
\boldsymbol{F}_{11} &= \boldsymbol{I}_{n_T} - \left(\bigoplus_{k=1}^{L}\mathbf{1}_{n_k}\right)\boldsymbol{E}^{-1}\left(\bigoplus_{k=1}^{L}\mathbf{1}_{n_k}^T\right) \\
&= \boldsymbol{I}_{n_T} - \bigoplus_{k=1}^{L}(\lambda+N_k)^{-1}\boldsymbol{J}_{n_k} - \lambda\left[\sum_{k=1}^{L}N_k(\lambda+N_k)^{-1}\right]^{-1} \\
&\times \quad [(\lambda+N_1)^{-1}\mathbf{1}_{n_1}^T,\ldots,(\lambda+N_L)^{-1}\mathbf{1}_{n_L}^T]^T. \qquad (5.34)
\end{aligned}$$

Hence,

$$\begin{aligned}
\boldsymbol{F}_{11.2} &= \boldsymbol{F}_{11} - \boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21} \\
&= \boldsymbol{I}_{n_T} - \bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\boldsymbol{J}_{n_k} - \lambda^{-1}\left[\sum_{k=1}^{L}(1-B_k)\right]^{-1} \\
&\times \quad [B_1\mathbf{1}_{n_1}^T,\ldots,B_L\mathbf{1}_{n_L}^T]^T[B_1\mathbf{1}_{n_1}^T,\ldots,B_L\mathbf{1}_{n_L}^T]. \qquad (5.35)
\end{aligned}$$

Hence,

$$\begin{aligned}
\boldsymbol{z}^T\boldsymbol{F}_{11.2}\boldsymbol{z} &= \sum_{k=1}^{L}\sum_{i\in str_k} z_i^2 - \sum_{k=1}^{L}(\lambda+n_k)^{-1}(n_k\bar{z}_k)^2 \\
&- \lambda\left[\sum_{k=1}^{L}(1-B_k)\right]^{-1}\left\{\sum_{k=1}^{L}(1-B_k)\bar{z}_k\right\}^2 \\
&= \sum_{k=1}^{L}\sum_{i\in str_k}(z_i-\bar{z}_k)^2 + \sum_{k=1}^{L}(n_k - n_k^2(\lambda+n_k)^{-1})\bar{z}_k^2 \\
&- \lambda\left[\sum_{k=1}^{L}(1-B_k)\right]^{-1}\left\{\sum_{k=1}^{L}(1-B_k)\bar{z}_k\right\}^2 \\
&= \sum_{k=1}^{L}\sum_{i\in str_k}(z_i-\bar{z}_k)^2 + \lambda\left\{\sum_{k=1}^{L}(1-B_k)(\bar{z}_k-\bar{z})^2\right\} \\
&= Q(\boldsymbol{z}). \qquad (5.36)
\end{aligned}$$

This completes the proof of the theorem. □

Using the above theorem, the hierarchical Bayes estimator of $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_L)^T$ is given by $\tilde{\boldsymbol{\mu}}^{HB} = (\tilde{\mu}_1^{HB},\ldots,\tilde{\mu}_L^{HB})^T$ where

$$\tilde{\mu}_k^{HB} = N_k^{-1}[n_k\bar{z}_k + (N_k - n_k)((1-B_k)\bar{z}_k + B_k\bar{z})]$$

$$= \bar{z}_k - f_k B_k(\bar{z}_k - \bar{z}), \tag{5.37}$$

$k = 1, \ldots, L$, which agrees with the EB estimator derived in the previous chapter for known $\lambda = \sigma^2/\tau^2$. However, using the posterior variance–covariance matrix derived in Theorem 5.2, the Bayes risk of the HB estimator derived in (5.37) differs from that of the EB estimator of the previous chapter when λ is known under the average squared error loss given in (4.22). The HB estimator as given in (5.37) is also derivable as the best unbiased predictor of $\boldsymbol{\mu}$ (best linear unbiased predictor without the normality assumption) when λ is known. This will follow as a special case of a more general result to be derived in the next section in the presence of auxiliary information.

Next we consider the case when σ^2 and τ^2 are both unknown so that λ is also unknown. We reparametrize $\sigma^2 = r^{-1}$ and $\tau^2 = (Mr)^{-1}$. We assign the gamma$(a/2, b/2)$ distribution to σ^2 and the gamma$(c/2, d/2)$ distribution to τ^2. To find the predictive distribution of $\boldsymbol{y}(s')$ given $\boldsymbol{z}$, first simplify using Exercise 2.4, p. 32, of Rao (1973),

$$\begin{aligned} |\boldsymbol{E}| &= |(\boldsymbol{D} + \lambda \boldsymbol{I}_L) - \lambda L^{-1} \boldsymbol{J}_L| \\ &= |\boldsymbol{D} + \lambda \boldsymbol{I}_L|(1 - \lambda L^{-1} \mathbf{1}_L^T (\boldsymbol{D} + \lambda I_L)^{-1} \mathbf{1}_L) \\ &= \prod_{k=1}^{L} (\lambda + N_k) \sum_{k=1}^{L} N_k (\lambda + N_k)^{-1}. \end{aligned} \tag{5.38}$$

Now re-express (5.18) as the joint (improper) pdf of $\boldsymbol{z}$, $\boldsymbol{y}(\bar{s})$ and λ, and write it as

$$\begin{aligned} f(\boldsymbol{z}, \boldsymbol{y}(\bar{s}), \lambda) \propto \prod_{k=1}^{L} \{\lambda(\lambda + N_k)^{-1}\}^{1/2} \left(\sum_{k=1}^{L} N_k (\lambda + N_k)^{-1} \right)^{-1/2} \\ \times (h(\boldsymbol{z}, \boldsymbol{y}(s)))^{-\frac{1}{2}(N_T + b + d - 1)}, \end{aligned} \tag{5.39}$$

where

$$h(\boldsymbol{z}, \boldsymbol{y}(s)) = a + c\lambda + \boldsymbol{z}^T \boldsymbol{z} + \boldsymbol{y}^T(\bar{s}) \boldsymbol{y}(\bar{s}) - \boldsymbol{w}^T \boldsymbol{E}^{-1} \boldsymbol{w}.$$

Now use (5.19)–(5.23), and integrate with respect to $\boldsymbol{y}(\bar{s})$ to get

$$\begin{aligned} f(\lambda|\boldsymbol{z}) &\propto f(\boldsymbol{z}, \lambda) \\ &\propto \left[\prod_{k=1}^{L} (\lambda(\lambda + N_k)^{-1}) \right]^{1/2} \left(\sum_{k=1}^{L} N_k (\lambda + N_k)^{-1} \right)^{-1/2} \end{aligned}$$

$$\times (a + c\lambda + \boldsymbol{z}^T \boldsymbol{F}_{11.2} \boldsymbol{z})^{-(N_T+b+d-1)/2}. \tag{5.40}$$

Using Exercise 2.4, p. 32, of Rao (1973) once again, one gets

$$\begin{aligned} |\boldsymbol{F}_{22}| &= |\boldsymbol{I}_{n_T}||\boldsymbol{E} - Diag(N_1 - n_1, \ldots, N_L - n_L)| \div |\boldsymbol{E}| \\ &= |Diag(\lambda + n_1, \ldots, \lambda + n_L) - \lambda L^{-1} \boldsymbol{J}_L| \\ &\div |Diag(\lambda + N_1, \ldots, \lambda + N_L) - \lambda L^{-1} \boldsymbol{J}_L| \\ &= \left[\prod_{k=1}^{L} (\lambda + n_k)\right] \left(\sum_{k=1}^{L} (1 - B_k)\right) \\ &\div \left[\prod_{k=1}^{L} (\lambda + N_k) \left(\sum_{k=1}^{L} N_k (\lambda + N_k)^{-1}\right)\right]. \end{aligned} \tag{5.41}$$

Combining (5.40) and (5.41), one gets

$$\begin{aligned} f(\lambda|\boldsymbol{z}) &\propto \prod B_k^{1/2} \left(\sum_{k=1}^{L} (1 - B_k)\right)^{-1/2} \\ &\times (a + c\lambda + \boldsymbol{z}^T \boldsymbol{F}_{11.2} \boldsymbol{z})^{-\frac{1}{2}(n_T+b+d-1)}. \end{aligned} \tag{5.42}$$

The above calculations lead to the following theorem.

Theorem 5.3 *Under the hierarchical model given in (I) and (II) of Section 4.3 and (III) of the present section. the predictive distribution of* $\boldsymbol{y}(\bar{s})$ *given* $\boldsymbol{z}$ *is as follows :*

(i) conditional on λ *and* $\boldsymbol{z}$*, the joint distribution of* $\boldsymbol{y}(\bar{s})$ *will be multivariate-t with location-vector, scale-matrix and degrees of freedom as given in Theorem 5.2;*

(ii) the conditional pdf of λ *given* $\boldsymbol{z}$ *is given in (5.42).*

Based on the above theorem, the HB predictor of $\boldsymbol{\mu}$ is given by $\hat{\boldsymbol{\mu}}^{HB} = (\hat{\mu}_1^{HB}, \ldots, \hat{\mu}_L^{HB})^T$, where

$$\hat{\mu}_k^{HB} = \bar{z}_k - f_k E[B_k(\bar{z}_k - \bar{z})|\boldsymbol{z}], \tag{5.43}$$

$k = 1, \ldots, L$. Also, the associated posterior variances and covariances are given by

$$\begin{aligned} V(\hat{\mu}_k|\boldsymbol{z}) &= V\left[N_k^{-1}\left(\sum_{i \in str_k} z_i + \sum_{j \in str_k, j \in \bar{s}} y_j\right)|\boldsymbol{z}\right] \\ &= N_k^{-2} V\left(\sum_{j \in str_k, j \in \bar{s}} y_j|\boldsymbol{z}\right) \end{aligned}$$

$$
\begin{aligned}
&= N_k^{-2}\left[V\left\{E\left(\sum_{j\in str_k, j\in\bar{s}} y_j|z\right)|z\right\}\right. \qquad (5.44)\\
&+ \left. E\left\{V\left(\sum_{j\in str_k, j\in\bar{s}} y_j|z\right)|z\right\}\right]\\
&= N_k^{-2}[V\{(N_k-n_k)((1-B_k)\bar{z}_k+B_k\bar{z})|z\}\\
&+ E\{(n_T+b+d-3)^{-1}(a+c\lambda+Q(z))G_{kk}|z\}].
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
Cov(\hat{\mu}_k,\hat{\mu}_{k'}) &= N_k^{-1}N_{k'}^{-1}Cov\Big[\sum_{i\in str_k} z_i+\sum_{j\in str_k, j\in\bar{s}} y_j,\\
&\qquad \sum_{i\in str_{k'}} z_i+\sum_{j\in str_{k'}, j\in\bar{s}} y_j|z\Big]\\
&= f_k f_{k'} Cov[B_k(\bar{z}_k-\bar{z}), B_{k'}(\bar{z}_{k'}-\bar{z})|z]. \qquad (5.45)
\end{aligned}
$$

Evaluation of these variances and covariances typically require numerical integration.

5.4 Auxiliary information I

This section generalizes the Bayesian model of the previous section by incorporating auxiliary information. As in the previous section, we first provide the analysis when the variance ratio is known. Specifically, consider the following hierarchical model:

(I) conditional on $\boldsymbol{\theta}$, $R = r$, and $\boldsymbol{b}$, the y_i are mutually independent normal, and for units belonging to stratum k, $E(y_i|\boldsymbol{\theta}, r, \boldsymbol{b}) = \theta_k$, $V(y_i|\boldsymbol{\theta}, r, \boldsymbol{b}) = r^{-1}$;

(II) conditional on r, and $\boldsymbol{b}$, θ_k are independent normal with $E(\theta_k|r, \boldsymbol{b}) = \boldsymbol{x}_k^T\boldsymbol{b}$, $V(\theta_k|r, \boldsymbol{b}) = (\lambda r)^{-1}$;

(III) $\boldsymbol{b}$, and R are mutually independent with $\boldsymbol{b} \sim$ uniform(R^p), $R \sim$ gamma$(a/2, b/2)$.

Note that in the above formulation, λ, the ratio of the superpopulation variance to the prior variance is assumed to be known. This assumption will be dispensed with in the later part of this section. The following theorem provides the predictive distribution of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$. As in the previous section, the Kronecker sum

is denoted by $\bigoplus$. Also, recall the definitions of B_k, $k = 1, \ldots, L$ and $\boldsymbol{G}$ from the previous section. Moreover, let $\tilde{\boldsymbol{b}} = [\sum_{k=1}^{L}(1 - B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T]^{-1}[\sum_{k=1}^{L}(1 - B_k)\boldsymbol{x}_k\bar{z}_k]$.

Theorem 5.4 *Consider the hierarchical model given in (I)-(III) of this section. Then the predictive distribution of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$ is multivariate-t with location-vector*

$$[(1 - B_1)\bar{z}_1 + B_1\boldsymbol{x}_1^T\tilde{\boldsymbol{b}}, \ldots, (1 - B_L)\bar{z}_L + B_L\boldsymbol{x}_L^T\tilde{\boldsymbol{b}}]^T,$$

scale-matrix $(n_T + b - 1)^{-1}(a + Q(\boldsymbol{z}))\boldsymbol{G}$ and degrees of freedom $n_T + b - 1$, where

$$Q(\boldsymbol{z}) = \sum_{k=1}^{L}\sum_{i\in str_k, i\in s}(y_i - \bar{z}_k)^2 + \lambda\sum_{k=1}^{L}(1 - B_k)(\bar{z}_k - \boldsymbol{x}_k^T\tilde{\boldsymbol{b}})^2.$$

Proof. Write the joint pdf of $\boldsymbol{z}$, $\boldsymbol{y}(\bar{s})$, $\boldsymbol{\theta}$, $\boldsymbol{b}$, and R as

$$\begin{aligned} &f(\boldsymbol{z}, \boldsymbol{y}(\bar{s}), \boldsymbol{\theta}, \boldsymbol{b}, r) \propto r^{N_T/2} \\ &\times\exp\left[-(r/2)\sum_{k=1}^{L}\left\{\sum_{i\in str_k, i\in s}(y_i - \theta_k)^2 + \sum_{j\in str_k, j\in\bar{s}}(y_j - \theta_k)^2\right\}\right] \\ &\times(\lambda r)^{L/2}\exp\left[-(\lambda r/2)\sum_{k=1}^{L}(\theta_k - \boldsymbol{x}_k^T\boldsymbol{b})^2\right]\exp(-ar/2)r^{b/2-1}. \end{aligned} \tag{5.46}$$

Write $\boldsymbol{X}^T = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$ and $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. First integrating with respect to $\boldsymbol{b}$, it follows that the joint pdf of $\boldsymbol{z}$, $\boldsymbol{y}(\bar{s})$, θ and R is

$$\begin{aligned} f(\boldsymbol{z}, \boldsymbol{y}(\bar{s}), \boldsymbol{\theta}, r) \quad \propto \quad & r^{(N_T+L+b-1)/2}\lambda^{(L-1)/2} \\ & \times \quad \exp\left[-(r/2)\left(\sum_{k=1}^{L}\left\{\sum_{i\in str_k, i\in s}(y_i - \theta_k)^2\right.\right.\right. \\ & + \quad \left.\left.\left.\sum_{j\in str_k, j\in\bar{s}}(y_j - \theta_k)^2\right\} + \lambda\boldsymbol{\theta}^T(I_L - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\theta}\right)\right]. \end{aligned} \tag{5.47}$$

Next simplify

$$\sum_{k=1}^{L}\left\{\sum_{i\in str_k, i\in s}(y_i - \theta_k)^2 + \sum_{j\in str_k, j\in\bar{s}}(y_j - \theta_k)^2\right\}$$

$$
\begin{aligned}
&+ \quad \lambda\boldsymbol{\theta}^T(I_{N_T} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{\theta} \\
&= \quad \boldsymbol{\theta}^T(\boldsymbol{D} + \lambda(\boldsymbol{I}_{N_T} - \boldsymbol{P}_{\boldsymbol{X}}))\boldsymbol{\theta} - 2\boldsymbol{\theta}^T((\bigoplus \mathbf{1}_{n_k})\boldsymbol{z} \\
&+ \quad (\bigoplus \mathbf{1}_{N_k - n_k})\boldsymbol{y}(\bar{s})) + \boldsymbol{z}^T\boldsymbol{z} + \boldsymbol{y}^T(\bar{s})\boldsymbol{y}(\bar{s}). \qquad (5.48)
\end{aligned}
$$

Write

$$
\boldsymbol{E} = \boldsymbol{D} + \lambda(\boldsymbol{I}_{N_T} - \boldsymbol{P}_{\boldsymbol{X}}). \qquad (5.49)
$$

As in the previous section, we find that the predictive density of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$ is multivariate-t with location-parameter $\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}\boldsymbol{y}$, scale-parameter $(n_T+b-1)^{-1}[a+\boldsymbol{z}^T(\boldsymbol{F}_{11}-\boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21})\boldsymbol{z}]\boldsymbol{F}_{22}^{-1}$ and degrees of freedom $n_T + b - 1$, where $\boldsymbol{F}_{11}, \boldsymbol{F}_{22}$ and $\boldsymbol{F}_{21}$ are defined in (5.20)–(5.22) with $\boldsymbol{E}$ redefined in 5.49. Thus, our calculations proceed as in the previous section with changes in the formulas occuring due to a more general $\boldsymbol{E}$. First find

$$
\begin{aligned}
\boldsymbol{F}_{22}^{-1} &= \boldsymbol{I}_{N_T - n_T} + \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k - n_k}\right) \\
&\times \quad (\boldsymbol{E} - Diag(N_1 - n_1, \ldots, N_L - n_L))^{-1} \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k - n_k}^T\right) \\
&= \boldsymbol{I}_{N_T - n_T} + \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k - n_k}\right) \\
&\times \quad (Diag(\lambda + n_1)^{-1}, \ldots, (\lambda + n_L)^{-1}) \\
&+ \quad Diag(B_1, \ldots, B_L) \\
&\times \quad \boldsymbol{X}(\boldsymbol{X}^T Diag(1 - B_1, \ldots, 1 - B_L)\boldsymbol{X})^{-1} \\
&\times \quad \boldsymbol{X}^T Diag((\lambda + n_1)^{-1}, \ldots, (\lambda + n_L)^{-1}) \left(\bigoplus_{k=1}^{L} \mathbf{1}_{N_k - n_k}^T\right) \\
&= \boldsymbol{I}_{N_T - n_T} + \bigoplus_{k=1}^{L} (\lambda + n_k)^{-1} \boldsymbol{J}_{N_k - n_k} \\
&+ \quad \lambda^{-1} \left(\bigoplus_{k=1}^{L} B_k \mathbf{1}_{N_k - n_k}\right) \boldsymbol{X} \\
&\times \quad \left(\sum_{k=1}^{L} (1 - B_k)\boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \boldsymbol{X}^T \left(\bigoplus_{k=1}^{L} B_k \mathbf{1}_{N_k - n_k}^T\right). \qquad (5.50)
\end{aligned}
$$

Next find

$$\begin{aligned}
\boldsymbol{E}^{-1} &= [Diag(\lambda+N_1,\ldots,\lambda+N_L)-\lambda\boldsymbol{P}_{\boldsymbol{X}}]^{-1} \\
&= Diag((\lambda+N_1)^{-1},\ldots,(\lambda+N_L)^{-1}) \\
&+ \lambda Diag((\lambda+N_1)^{-1},\ldots,(\lambda+N_L)^{-1})\boldsymbol{X} \\
&\times \left(\sum_{k=1}^{L} N_k(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1} \\
&\times \boldsymbol{X}^T Diag((\lambda+N_1)^{-1},\ldots,(\lambda+N_L)^{-1}). \qquad (5.51)
\end{aligned}$$

Hence,

$$\begin{aligned}
\boldsymbol{F}_{21} &= \bigoplus_{k=1}^{L}(\lambda+N_k)^{-1}\mathbf{1}_{N_k-n_k}\mathbf{1}_{n_k}^T \\
&+ \lambda\left(\bigoplus_{k=1}^{L}(\lambda+N_k)^{-1}\mathbf{1}_{N_k-n_k}\right)\boldsymbol{X} \\
&\times \left(\sum_{k=1}^{L} N_k(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1} \\
&\times \boldsymbol{X}^T\left(\bigoplus_{k=1}^{L}(\lambda+N_k\right)^{-1}\mathbf{1}_{n_k}^T. \qquad (5.52)
\end{aligned}$$

Hence,

$$\begin{aligned}
\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21} &= \bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{N_k-n_k}\mathbf{1}_{n_k}^T \\
&+ \left(\bigoplus_{k=1}^{L} B_k\mathbf{1}_{N_k-n_k}\right)\boldsymbol{X}\left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{X}^T \\
&\times \left(\bigoplus_{k=1}^{L}(\lambda+N_k)^{-1}(\lambda+n_k)^{-1}(N_k-n_k)\mathbf{1}_{n_k}^T\right) \\
&+ \lambda\left(\bigoplus_{k=1}^{L}(\lambda+N_k)^{-1}\mathbf{1}_{N_k-n_k}\right)\boldsymbol{X} \\
&\times \left(\sum_{k=1}^{L} N_k(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}
\end{aligned}$$

$$
\begin{aligned}
&\times \quad \boldsymbol{X}^T \left(\bigoplus_{k=1}^{L} (\lambda + N_k)^{-1} \mathbf{1}_{n_k}^T \right) \\
&+ \quad \lambda \left(\bigoplus_{k=1}^{L} \{(\lambda + n_k)^{-1} - (\lambda + N_k)^{-1}\} \mathbf{1}_{N_k - n_k} \right) \boldsymbol{X} \\
&\times \quad \left(\sum_{k=1}^{L} N_k (\lambda + N_k)^{-1} \boldsymbol{x}_k \boldsymbol{x}_k^T \right)^{-1} \\
&\times \quad \boldsymbol{X}^T \left(\bigoplus_{k=1}^{L} (\lambda + N_k)^{-1} \mathbf{1}_{n_k}^T \right) \\
&+ \quad \left(\bigoplus_{k=1}^{L} B_k \mathbf{1}_{N_k - n_k} \right) \boldsymbol{X} \left(\sum_{k=1}^{L} (1 - B_k) \boldsymbol{x}_k \boldsymbol{x}_k^T \right)^{-1} \boldsymbol{X}^T \\
&\times \quad \left(\bigoplus_{k=1}^{L} (\lambda + N_k)^{-1} (\lambda + n_k)^{-1} (N_k - n_k) \mathbf{1}_{N_k - n_k} \right) \boldsymbol{X} \\
&\times \quad \left(\sum_{k=1}^{L} N_k (\lambda + N_k)^{-1} \boldsymbol{x}_k \boldsymbol{x}_k^T \right)^{-1} \\
&\times \quad \boldsymbol{X}^T \left(\bigoplus_{k=1}^{L} (\lambda + N_k)^{-1} \mathbf{1}_{n_k}^T \right). \qquad (5.53)
\end{aligned}
$$

Write

$$
\begin{aligned}
&\boldsymbol{X}^T \left(\bigoplus_{k=1}^{L} (\lambda + N_k)^{-1} (\lambda + n_k)^{-1} (N_k - n_k) \right) \boldsymbol{X} \\
&= \quad (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \left\{ \bigoplus_{k=1}^{L} N_k (\lambda + N_k)^{-1} - n_k (\lambda + n_k)^{-1} \right\} \\
&\times \quad (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)^T \\
&= \quad \sum_{k=1}^{L} \{N_k (\lambda + N_k)^{-1} - n_k (\lambda + n_k)^{-1}\} \boldsymbol{x}_k \boldsymbol{x}_k^T. \qquad (5.54)
\end{aligned}
$$

This leads to

$$
\boldsymbol{F}_{22}^{-1} \boldsymbol{F}_{21} = \bigoplus_{k=1}^{L} (\lambda + n_k)^{-1} \mathbf{1}_{N_k - n_k} \mathbf{1}_{n_k}^T
$$

$$
\begin{aligned}
&+\left(\bigoplus_{k=1}^{L} B_k \mathbf{1}_{N_k-n_k}\right) \boldsymbol{X} \left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1} \boldsymbol{X}^T \\
&\times \left(\bigoplus_{k=1}^{L}\{(\lambda+n_k)^{-1}-(\lambda+N_k)^{-1}\}\mathbf{1}_{n_k}^T\right) \\
&+\lambda\left(\bigoplus_{k=1}^{L}(\lambda+N_k)^{-1}\mathbf{1}_{N_k-n_k}\right)\boldsymbol{X} \\
&\times \left(\sum_{k=1}^{L} N_k(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1} \boldsymbol{X}^T \left(\bigoplus_{k=1}^{L}(\lambda+N_k)^{-1}\mathbf{1}_{n_k}^T\right) \\
&+\lambda\left(\bigoplus_{k=1}^{L}\{(\lambda+n_k)^{-1}-(\lambda+N_k)^{-1}\}\mathbf{1}_{N_k-n_k}\right)\boldsymbol{X} \\
&\times \left(\sum_{k=1}^{L} N_k(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1} \boldsymbol{X}^T \left(\bigoplus_{k=1}^{L}(\lambda+N_k)^{-1}\mathbf{1}_{n_k}^T\right) \\
&+\left(\bigoplus_{k=1}^{L} B_k \mathbf{1}_{N_k-n_k}\right)\boldsymbol{X} \\
&\times \left\{\left(\sum_{k=1}^{L}\lambda(\lambda+n_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1} - \left(\sum_{k=1}^{L}\lambda(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\right\} \\
&\times \boldsymbol{X}^T\left(\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{n_k}^T\right).
\end{aligned}
$$

On simplification, one gets

$$
\begin{aligned}
\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21} &= \bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{N_k-n_k}\mathbf{1}_{n_k}^T + \left(\bigoplus_{k=1}^{L} B_k\mathbf{1}_{N_k-n_k}\right)\boldsymbol{X} \\
&\times \left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{X}^T\left(\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{n_k}^T\right).
\end{aligned}
\tag{5.55}
$$

Thus

$$
\begin{aligned}
\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}\boldsymbol{z} &= [(1-B_1)\bar{z}_1\mathbf{1}_{N_1-n_1}^T,\ldots,(1-B_L)\bar{z}_L\mathbf{1}_{N_L-n_L}^T]^T \\
&+ [B_1\boldsymbol{x}_1\mathbf{1}_{N_1-n_1}^T,\ldots,B_L\boldsymbol{x}_L\mathbf{1}_{N_L-n_L}^T]^T
\end{aligned}
$$

$$\times \left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}$$
$$\times (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L)[(1-B_1)\bar{z}_1,\ldots,(1-B_L)\bar{z}_L]^T$$
$$= [h_1\mathbf{1}_{N_1-n_1}^T,\ldots,h_L\mathbf{1}_{N_L-n_L}^T]^T, \tag{5.56}$$

where $h_k = (1-B_k)\bar{z}_k + B_k\boldsymbol{x}_k^T\tilde{\boldsymbol{b}}$, $k = 1,\ldots,L$. Also,

$$\begin{aligned}
\boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21} = {} & \left(\bigoplus_{k=1}^{L}\mathbf{1}_{n_k}\right)\Bigg[Diag((\lambda+N_1)^{-1},\ldots,(\lambda+N_L)^{-1}) \\
& + \lambda Diag((\lambda+N_1)^{-1},\ldots,(\lambda+N_L)^{-1})\boldsymbol{X} \\
& \times \left(\sum_{k=1}^{L}N_k(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{X}^T \\
& \times Diag((\lambda+N_1)^{-1},\ldots,(\lambda+N_L)^{-1})\Bigg] \\
& \times \left(\bigoplus_{k=1}^{L}\mathbf{1}_{N_k-n_k}^T\right) \\
& \times \left[\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{N_k-n_k}\mathbf{1}_{n_k}^T\right] \\
& + \left(\bigoplus_{k=1}^{L}B_k\mathbf{1}_{N_k-n_k}\right)\boldsymbol{X} \\
& \times \left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{X}^T \\
& \times \left(\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{n_k}^T\right).
\end{aligned}$$

After much algebraic simplification, one has

$$\begin{aligned}
\boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21} = {} & \bigoplus_{k=1}^{L}[(\lambda+n_k)^{-1}-(\lambda+N_k)^{-1}]\boldsymbol{J}_{n_k} \\
& + \bigoplus_{k=1}^{L}B_k\mathbf{1}_{n_k}
\end{aligned}$$

$$\times \quad \boldsymbol{X}\left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}\boldsymbol{X}^T$$
$$\times \quad \left(\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{n_k}^T\right)$$
$$- \quad \left(\bigoplus_{k=1}^{L}\lambda(\lambda+N_k)^{-1}\mathbf{1}_{n_k}\right)\boldsymbol{X}$$
$$\times \quad \left(\sum_{k=1}^{L}N_k(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}$$
$$\times \quad \boldsymbol{X}^T\left(\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{n_k}^T\right). \tag{5.57}$$

Also,

$$\boldsymbol{F}_{11} = \boldsymbol{I}_{N_T} - \bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\boldsymbol{J}_{n_k}$$
$$- \lambda\left(\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{n_k}\right)\boldsymbol{X}\left(\sum_{k=1}^{L}N_k(\lambda+N_k)^{-1}\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}$$
$$\times \boldsymbol{X}^T\left(\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{n_k}^T\right). \tag{5.58}$$

Combining (5.57) and (5.58), one gets

$$\boldsymbol{F}_{11.2} = \boldsymbol{F}_{11} - \boldsymbol{F}_{12}\boldsymbol{F}_{22}^{-1}\boldsymbol{F}_{21}$$
$$= \boldsymbol{I}_{N_T} - \bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\boldsymbol{J}_{n_k}$$
$$- \left(\bigoplus_{k=1}^{L}B_k\mathbf{1}_{n_k}\right)\boldsymbol{X}\left(\sum_{k=1}^{L}(1-B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)^{-1}$$
$$\times \boldsymbol{X}^T\left(\bigoplus_{k=1}^{L}(\lambda+n_k)^{-1}\mathbf{1}_{n_k}^T\right). \tag{5.59}$$

Hence,

$$Q(\boldsymbol{z}) = \boldsymbol{z}^T\boldsymbol{F}_{11.2}\boldsymbol{z}$$

$$
\begin{aligned}
&= \sum_{k=1}^{L} \sum_{i \in str_k, i \in s} (y_i - \bar{z}_k)^2 + \sum_{k=1}^{L} n_k \bar{z}_k^2 \\
&- \sum_{k=1}^{L} n_k^2 (\lambda + n_k)^{-1} \bar{z}_k^2 - \lambda \sum_{k=1}^{L} (1 - B_k) \bar{z}_k \boldsymbol{x}_k^T \tilde{\boldsymbol{b}} \\
&= \sum_{k=1}^{L} \sum_{i \in str_k, i \in s} (y_i - \bar{z}_k)^2 + \lambda \sum_{k=1}^{L} (1 - B_k) \bar{z}_k^2 \\
&- \lambda \sum_{k=1}^{L} (1 - B_k) \bar{z}_k \boldsymbol{x}_k^T \tilde{\boldsymbol{b}} \\
&= \sum_{k=1}^{L} \sum_{i \in str_k, i \in s} (y_i - \bar{z}_k)^2 + \lambda \sum_{k=1}^{L} (1 - B_k) (\bar{z}_k - \boldsymbol{x}_k^T \tilde{\boldsymbol{b}})^2,
\end{aligned}
\tag{5.60}
$$

since

$$
\tilde{\boldsymbol{b}} = \left(\sum_{k=1}^{L} (1 - B_k) \boldsymbol{x}_k \boldsymbol{x}_k^T \right)^{-1} \sum_{k=1}^{L} (1 - B_k) \bar{z}_k \boldsymbol{x}_k,
$$

and

$$
\begin{aligned}
&\sum_{k=1}^{L} (1 - B_k) \bar{z}_k \tilde{\boldsymbol{b}}^T \boldsymbol{x}_k \boldsymbol{x}_k^T \tilde{\boldsymbol{b}} \\
&= \sum_{k=1}^{L} (1 - B_k) \bar{z}_k (\tilde{\boldsymbol{b}}^T \boldsymbol{x}_k) = \tilde{\boldsymbol{b}}^T \sum_{k=1}^{L} (1 - B_k) \bar{z}_k \boldsymbol{x}_k.
\end{aligned}
$$

This completes the proof of the theorem. □

It follows from the theorem that the Bayes predictor of $\mu_k = N_k^{-1} \sum_{i \in str_k} y_i$ is given by

$$
\begin{aligned}
E[\mu_k | \boldsymbol{z}] &= N_k^{-1} [n_k \bar{z}_k + (N_k - n_k)((1 - B_k) \bar{z}_k + B_k \boldsymbol{x}_k^T \tilde{\boldsymbol{b}})] \\
&= \bar{z}_k - f_k B_k (\bar{z}_k - \boldsymbol{x}_k^T \tilde{\boldsymbol{b}}).
\end{aligned}
\tag{5.61}
$$

Next we prove some frequentist optimality properties of these HB predictors. The next theorem shows that the HB predictor $\tilde{\boldsymbol{\mu}}^{HB}$ of $\boldsymbol{\mu}$ is its best unbiased predictor under some mixed-effects model. More specifically, it is assumed that for $i \in str_k$,

$$
y_i = \boldsymbol{x}_k^T \boldsymbol{b} + v_k + e_i,
\tag{5.62}
$$

where the v_k and the e_i are mutually independent with the v_k iid $N(0, \tau^2)$, while the e_i iid $N(0, \sigma^2)$. We now treat $\boldsymbol{b}$, σ^2 and τ^2 as the unknown parameters of the given model, but $\lambda = \sigma^2/\tau^2$ is known. Also, in what follows, we denote by E, V and Cov expectations, variances and covariances under the model given in (5.62). We need also the following definition of an unbiased predictor of $\boldsymbol{\mu}$.

Definition 1. A predictor $\boldsymbol{\delta}(\boldsymbol{z})$ of $\boldsymbol{\mu}$ is said to be **unbiased** if

$$E[\boldsymbol{\delta}(\boldsymbol{z}) - \boldsymbol{\mu}] = \mathbf{0} \text{ for all } \boldsymbol{b}, \sigma^2 \text{ and } \tau^2. \tag{5.63}$$

Also, $\boldsymbol{\delta}_0(\boldsymbol{z})$ is said to be a **best unbiased predictor** (BUP) of $\boldsymbol{\mu}$ if $\boldsymbol{\delta}_0(\boldsymbol{z})$ is an unbiased predictor of $\boldsymbol{\mu}$ and

$$E[(\boldsymbol{\delta}_0(\boldsymbol{z}) - \boldsymbol{\mu})(\boldsymbol{\delta}_0(\boldsymbol{z}) - \boldsymbol{\mu})^T] \leq E[(\boldsymbol{\delta}(\boldsymbol{z}) - \boldsymbol{\mu})(\boldsymbol{\delta}(\boldsymbol{z}) - \boldsymbol{\mu})^T]$$

for all $\boldsymbol{b}$, σ^2 and τ^2 for every unbiased predictor $\boldsymbol{\delta}(\boldsymbol{z})$ of $\boldsymbol{\mu}$. We shall prove that $\tilde{\boldsymbol{\mu}}^{HB}$ is the BUP of $\boldsymbol{\mu}$ under the model given in (5.62). This will be accomplished through several steps.

First note that our target is to minimize $E[(\boldsymbol{\delta}(\boldsymbol{z})-\boldsymbol{\mu})(\boldsymbol{\delta}(\boldsymbol{z})-\boldsymbol{\mu})^T]$ with respect to $\boldsymbol{\delta}$. Use the identity

$$\begin{aligned} &E[(\boldsymbol{\delta}(\boldsymbol{z}) - \boldsymbol{\mu})(\boldsymbol{\delta}(\boldsymbol{z}) - \boldsymbol{\mu})^T] \\ &= E[(\boldsymbol{\delta}(\boldsymbol{z}) - E(\boldsymbol{\mu}|\boldsymbol{z}))(\boldsymbol{\delta}(\boldsymbol{z}) - E(\boldsymbol{\mu}|\boldsymbol{z}))^T] + V(\boldsymbol{\mu}|\boldsymbol{z}). \end{aligned} \tag{5.64}$$

Thus, it suffices to minimize the first term in the right hand side of (5.64) with respect to $\boldsymbol{\delta}$. Note that under the model given in (5.62),

$$\begin{aligned} E(\boldsymbol{\mu}|\boldsymbol{z}) &= (\bar{z}_1 - f_1 B_1(\bar{z}_1 - \boldsymbol{x}_1^T \boldsymbol{b}), \ldots, \bar{z}_L - f_L B_L(\bar{z}_L - \boldsymbol{x}_L^T \boldsymbol{b}))^T \\ &= (\tilde{\mu}_1(\bar{z}_1), \ldots, \tilde{\mu}_L(\bar{z}_L))^T = \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}) \text{ (say)}. \end{aligned} \tag{5.65}$$

We shall find componentwise BUP's for $\tilde{\mu}_k(\bar{z}_k)$, $k = 1, \ldots, L$, and argue that the vector consisting of the componentwise BUP's is the BUP of $\tilde{\boldsymbol{\mu}}_{\boldsymbol{z}}$.

This is achieved through the use of the following lemma.

Lemma 5.4.1 *An unbiased predictor $\delta_{0k}(\boldsymbol{z})$ of $\tilde{\mu}_k(\bar{z}_k)$ is a BUP of $\tilde{\mu}_k(\bar{z}_k)$ if and only if $\delta_{0k}(\boldsymbol{z}) - \tilde{\mu}_k(\bar{z}_k)$ is uncorrelated with every $h(\boldsymbol{z})$ such that $E(h(\boldsymbol{z})) = 0$ for all $\boldsymbol{b}$, σ^2 and τ^2.*

Suppose for the moment that the lemma is true. Write $\boldsymbol{\delta}_0(\boldsymbol{z}) = (\delta_{01}(\boldsymbol{z}), \ldots, \delta_{0L}(\boldsymbol{z}))^T$. Let $\boldsymbol{\delta}(\boldsymbol{z}) = (\delta_1(\boldsymbol{z}), \ldots, \delta_L(\boldsymbol{z}))^T$ be also an unbiased predictor of $\tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}})$. We need to show that

$$\begin{aligned} &E[(\boldsymbol{\delta}_0(\boldsymbol{z}) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}))(\boldsymbol{\delta}_0(\boldsymbol{z}) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}))^T] \\ &\leq E[(\boldsymbol{\delta}(\boldsymbol{z}) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}))(\boldsymbol{\delta}(\boldsymbol{z}) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}))^T], \end{aligned} \tag{5.66}$$

where for any two symmetric matrices $\boldsymbol{F}$ and $\boldsymbol{G}$, we say that $\boldsymbol{F} \leq \boldsymbol{G}$ if $\boldsymbol{G} - \boldsymbol{F}$ is non-negative definite. Now observe that for every $\boldsymbol{a} = (a_1, \ldots, a_L)^T$,

$$\begin{aligned} & \boldsymbol{a}^T E[(\boldsymbol{\delta}(\boldsymbol{z}) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}))(\boldsymbol{\delta}(\boldsymbol{z}) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}))^T]\boldsymbol{a} \\ - \; & \boldsymbol{a}^T E[(\boldsymbol{\delta}_0(\boldsymbol{z}) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}))(\boldsymbol{\delta}_0(\boldsymbol{z}) - \tilde{\boldsymbol{\mu}}(\bar{\boldsymbol{z}}))^T]\boldsymbol{a} \\ = \; & E\left[\sum_{k=1}^{L} a_k \delta_k(\boldsymbol{z}) - \sum_{k=1}^{L} a_k \tilde{\mu}_k(\bar{\boldsymbol{z}})\right]^2 \\ - \; & E\left[\sum_{k=1}^{L} a_k \delta_{0k}(\boldsymbol{z}) - \sum_{k=1}^{L} a_k \tilde{\mu}_k(\bar{\boldsymbol{z}})\right]^2 \geq 0, \end{aligned} \tag{5.67}$$

since for every $h(\boldsymbol{z})$ with $E[h(\boldsymbol{z})] = 0$ for all $\boldsymbol{b}$, σ^2 and τ^2,

$$Cov\left(\sum_{k=1}^{L} a_k \delta_{0k}(\boldsymbol{z}), h(\boldsymbol{z})\right) = 0,$$

so that $\sum_{k=1}^{L} a_k \delta_{0k}(\boldsymbol{z})$ is the BUP of $\sum_{k=1}^{L} a_k \tilde{\mu}_k(\bar{\boldsymbol{z}})$. Next note that

$$\bar{z}_k - f_k B_k(\bar{z}_k - \bar{\bar{z}}) - \tilde{\mu}_k(\bar{z}_k) = \boldsymbol{x}_k^T(\boldsymbol{b} - \tilde{\boldsymbol{b}}).$$

We are now in a position to prove the following theorem.

Theorem 5.5 *Under the model given in (5.62), $\tilde{\mu}^{HB}$ is the BUP of μ.*

Proof. In view of (5.64), (5.65), Lemma 5.4.1 and the discussion following the lemma, it suffices to show that

$$Cov(\boldsymbol{x}_k^T(\tilde{\boldsymbol{b}} - \boldsymbol{b}), h(\boldsymbol{z})) = 0 \tag{5.68}$$

for every $h(\boldsymbol{z})$ with $E[h(\boldsymbol{z})] = 0$ for all $\boldsymbol{b}$, σ^2 and τ^2. Equivalently, we need to show that $E[h(\boldsymbol{z})\boldsymbol{x}_k^T(\tilde{\boldsymbol{b}} - \boldsymbol{b})] = 0$ for all $\boldsymbol{b}$, σ^2 and τ^2. To prove this, first note that if $\boldsymbol{z}_k$ denotes the sampled z_i in stratum k, then

$$\boldsymbol{z}_k \sim N((\boldsymbol{x}_k^T\boldsymbol{b})\mathbf{1}_{n_k}, \sigma^2\boldsymbol{I}_{n_k} + \tau^2\boldsymbol{J}_{n_k}). \tag{5.69}$$

Hence, $E(h(\boldsymbol{z})) = 0$ can be written equivalently as

$$\begin{aligned} & \int \cdots \int h(\boldsymbol{z}) \prod |\sigma^2\boldsymbol{I}_{n_k} + \tau^2\boldsymbol{J}_{n_k}|^{-1/2}(2\pi)^{-n_T/2} \\ \times \; & \exp\left[-\sum_{k=1}^{L}(\boldsymbol{z}_k - (\boldsymbol{x}_k^T\boldsymbol{b})\mathbf{1}_{n_k})^T(\sigma^2\boldsymbol{I}_{n_k} + \tau^2\boldsymbol{J}_{n_k})^{-1}\right. \\ \times \; & \left.(\boldsymbol{z}_k - (\boldsymbol{x}_k^T\boldsymbol{b})\mathbf{1}_{n_k})/2\right] d\boldsymbol{z} = 0. \end{aligned} \tag{5.70}$$

Using the standard matrix inversion formula,

$$\begin{aligned}(\sigma^2 \boldsymbol{I}_{n_k} + \tau^2 \boldsymbol{J}_{n_k})^{-1} &= \sigma^{-2}[\boldsymbol{I}_{n_k} - \tau^2(\sigma^2 + n_k\tau^2)^{-1}\boldsymbol{J}_{n_k}] \\ &= \sigma^{-2}[\boldsymbol{I}_{n_k} - (\lambda + n_k)^{-1}\boldsymbol{J}_{n_k}]. \qquad (5.71)\end{aligned}$$

This leads to the identity

$$\begin{aligned}&\sum_{k=1}^{L}(\boldsymbol{z}_k - (\boldsymbol{x}_k^T\boldsymbol{b})\mathbf{1}_{n_k})^T[\sigma^2\boldsymbol{I}_{n_k} + \tau^2\boldsymbol{J}_{n_k}]^{-1}(\boldsymbol{z}_k - (\boldsymbol{x}_k^T\boldsymbol{b})\mathbf{1}_{n_k}) \\ &= \sigma^{-2}\left[\sum_{k=1}^{L}||\boldsymbol{z}_k - (\boldsymbol{x}_k^T\boldsymbol{b})\mathbf{1}_{n_k}||^2 - \sum_{k=1}^{L}(\lambda + n_k)^{-1}n_k^2(\bar{z}_k - \boldsymbol{x}_k^T\boldsymbol{b})^2\right] \\ &= \sigma^{-2}\left[\sum_{k=1}^{L}\sum_{i\in str_k}(z_i - \bar{z}_k)^2 + \sum_{k=1}^{L}n_k B_k(\bar{z}_k - \boldsymbol{x}_k^T\boldsymbol{b})^2\right] \\ &= \sigma^{-2}\sum_{k=1}^{L}\sum_{i\in str_k}(z_i - \bar{z}_k)^2 + \tau^{-2}\sum_{k=1}^{L}(1 - B_k)(\bar{z}_k - \boldsymbol{x}_k^T\boldsymbol{b})^2. \qquad (5.72)\end{aligned}$$

Using (5.72) and differentiating both sides of (5.70) with respect to $\boldsymbol{b}$, one gets

$$\begin{aligned}&\int\cdots\int h(\boldsymbol{z})\tau^{-2}\sum_{k=1}^{L}(1 - B_k)(\bar{z}_k - \boldsymbol{x}_k^T\boldsymbol{b})\boldsymbol{x}_k \\ &\quad\times \prod_{k=1}^{L}|\sigma^2\boldsymbol{I}_{n_k} + \tau^2\boldsymbol{J}_{n_k}|^{-1/2}(2\pi)^{-n_T/2} \\ &\quad\times \exp\left[-\frac{1}{2}\sum_{k=1}^{L}(\boldsymbol{z}_k - (\boldsymbol{x}_k^T\boldsymbol{b})\mathbf{1}_{n_k})^T(\sigma^2\boldsymbol{I}_{n_k} + \tau^2\boldsymbol{J}_{n_k})^{-1}\right. \\ &\quad\times \left.(\boldsymbol{z}_k - (\boldsymbol{x}_k^T\boldsymbol{b})\mathbf{1}_{n_k}/2\right] = \mathbf{0}, \qquad (5.73)\end{aligned}$$

which can be alternatively written as

$$E\left[h(\boldsymbol{z})\left\{\sum_{k=1}^{L}(1 - B_k)\bar{z}_k\boldsymbol{x}_k - \left(\sum_{k=1}^{L}(1 - B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)\boldsymbol{b}\right\}\right] = \mathbf{0},$$

that is

$$E\left[h(\boldsymbol{z})\left(\sum_{k=1}^{L}(1 - B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)(\tilde{\boldsymbol{b}} - \boldsymbol{b})\right] = \mathbf{0}$$

or

$$E[h(z)(\tilde{b} - b)] = 0, \tag{5.74}$$

since $\sum_{k=1}^{L}(1 - B_k)x_k x_k^T$ is nonstochastic and nonsingular. Premultiplication of (5.74) by x_k^T then leads to (5.69). This completes the proof of Theorem 5.5. □

It remains to prove Lemma 5.4.1. The proof follows the arguments of Lehmann and Scheffe (1950).

Proof. To prove the if part, note that for any unbiased predictor $\delta_k(z)$ of $\tilde{\mu}_k(\bar{z}_k)$,

$$\begin{aligned} & E[\delta_k(z) - \tilde{\mu}_k(\bar{z}_k)]^2 \\ &= E[\delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k)]^2 + E[\delta_k(z) - \delta_{0k}(z)]^2 \\ &+ 2E[\delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k)][\delta_k(z) - \delta_{0k}(z)]. \end{aligned} \tag{5.75}$$

But since $E[\delta_k(z) - \delta_{0k}(z)] = 0$, using the condition of the theorem,

$$\begin{aligned} & 2E[\delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k)][\delta_k(z) - \delta_{0k}(z)] \\ &= 2Cov(\delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k), \delta_k(z) - \delta_{0k}(z)) \\ &= 0. \end{aligned} \tag{5.76}$$

To prove the only if part, if $E[h(z)] = 0$ for any constant η, $\delta_{0k}(z) + \eta h(z)$ is also an unbiased predictor of $\tilde{\mu}_k(\bar{z}_k)$. This leads to the inequality

$$E[\delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k)]^2 \leq E[\delta_{0k}(z) + \eta h(z) - \tilde{\mu}_k(\bar{z}_k)]^2,$$

which is equivalent to

$$\eta[\eta V(h(z)) + 2Cov(h(z), \delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k))] \geq 0. \tag{5.77}$$

Suppose now $Cov(h(z), \delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k)) \neq 0$. If it is positive for some b_0, σ_0^2 and τ_0^2, then choose

$$\eta \in (-2\,Cov_{b_0,\sigma_0^2,\tau_0^2}(h(z), \delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k))/V(h(z)), 0),$$

to get a contradiction to (5.77). If this covariance is negative, then choose

$$\eta \in (0, -2Cov_{b_0,\sigma_0^2,\tau_0^2}(h(z), \delta_{0k}(z) - \tilde{\mu}_k(\bar{z}_k))/V(h(z)))$$

to get a contradiction to (5.77). This proves Lemma 5.4.1. □

Next we dispense with the normality assumption in (5.62), but retain the assumption of independence and finite second moments. We shall show that $\tilde{\mu}^{HB}$ is the best linear unbiased predictor

(BLUP) of μ, that is for every unbiased predictor $\delta(z)$ of μ which is linear in z,

$$E[(\tilde{\mu}^{HB} - \mu)(\tilde{\mu}^{HB} - \mu)^T \leq E[(\delta(z) - \mu)(\delta(z) - \mu)^T]. \quad (5.78)$$

Arguing as before, it suffices to show that $\tilde{\mu}_k^{HB}$ is the BLUP of μ_k for each $k = 1, \ldots, L$. Also, we note that for any linear unbiased predictor (LUP) $\delta_k(z) = \sum_{l=1}^{L} \sum_{i \in str_k} a_{kli} z_i$ of μ_k, under the model given in (5.62),

$$\begin{aligned}
E[\delta_k(z) - \mu_k]^2 &= V[\delta_k(z) - \mu_k] \\
&= V[\delta_k(z)] + V(\mu_k) - 2\,Cov(\delta_k(z), \mu_k) \\
&= (\sigma^2 + \tau^2) \sum_{l=1}^{L} \sum_{i \in str_l} a_{kli}^2 \\
&+ \tau^2 \sum_{l=1}^{L} \sum_{i \neq i', i \in s, i' \in s} \sum_{i \in str_l, i' \in str_l} a_{kli} a_{kli'} \\
&+ V(\mu_k) \\
&- 2(n_k/N_k) \sum_{l=1}^{L} (\sigma^2 n_l^{-1} + \tau^2) \sum_{i \in str_l, i \in s} a_{kli} \\
&= \sigma^2 \sum_{l=1}^{L} \left(\sum_{i \in s\, i \in str_l} a_{kli} \right)^2 + V(\mu_k) \\
&- 2 n_k N_k^{-1} (\sigma^2 n_k^{-1} + \tau^2) \sum_{i \in str_k, i \in s} a_{kki}. \quad (5.79)
\end{aligned}$$

It is clear from (5.79) that subject to $\sum_{i \in str_l, i \in s} a_{kli}$ being fixed, for all $l = 1, \ldots, L$, $\sum_{i \in str_l, i \in s} a_{kli}^2$, and accordingly $E[\delta_k(z) - \mu_k]^2$ is minimized when a_{kli} is the same for all i in stratum l. Thus, we can restrict ourselves to linear unbiased predictors of the form $\sum_{l=1}^{L} a_{kl} \bar{z}_l$ for μ_k. Arguing similarly as Lemma 5.4.1, it suffices now to prove the following lemma.

Lemma 5.4.2 *Under the model given in (5.62) without the normality assumption,*

$$Cov\left(\tilde{\mu}_k^{HB} - \mu_k, \sum_{l=1}^{L} c_{kl} \bar{z}_l \right) = 0, \quad (5.80)$$

for all $\mathbf{b}$, σ^2 *and* τ^2, *where* $E[\sum_{l=1}^{L} c_{kl} \bar{z}_l] = 0$ *for all* $\mathbf{b}$, σ^2 *and* τ^2.

Proof. Since $E[\sum_{l=1}^{L} c_{kl}\bar{z}_l] = \sum_{l=1}^{L} c_{kl}\boldsymbol{x}_l^T\boldsymbol{b}$, by hypothesis of the theorem, $\sum_{l=1}^{L} c_{kl}\boldsymbol{x}_l = \mathbf{0}$. To prove (5.80), first write

$$\tilde{\mu}_k^{HB} - \mu_k = f_k[(1 - B_k)\bar{z}_k + B_k\bar{\boldsymbol{x}}_k^T\tilde{\boldsymbol{b}} - \bar{y}_k(\bar{s})], \tag{5.81}$$

where $\bar{y}_k(\bar{s})$ is the average of the unsampled units for the kth stratum. From (5.80) and (5.81) one gets

$$\begin{aligned} &Cov\left(\tilde{\mu}_k^{HB} - \mu_k, \sum_{l=1}^{L} c_{kl}\bar{z}_l\right) \\ &= f_k\, Cov\left[(1 - B_k)\bar{z}_k + B_k\bar{\boldsymbol{x}}_k^T\tilde{\boldsymbol{b}} - \bar{y}_k(\bar{s}), \sum_{l=1}^{L} c_{kl}\bar{z}_l\right] \\ &= f_k\,[c_{kk}\{(1 - B_k)V(\bar{z}_k) - Cov(\bar{y}_k(\bar{s}), \bar{z}_k)\} \\ &+ B_k\, Cov\left(\boldsymbol{x}_k^T\tilde{\boldsymbol{b}}, \sum_{l=1}^{L} c_{kl}\bar{z}_l\right). \end{aligned} \tag{5.82}$$

Now observe that

$$V(\bar{z}_k) = \sigma^2 n_k^{-1} + \tau^2 = \sigma^2 M^{-1}(1 - B_k)^{-1} = \tau^2(1 - B_k)^{-1}. \tag{5.83}$$

Also, $Cov(\bar{y}_k(\bar{s}), \bar{z}_k) = V(v_k) = \tau^2$. Hence,

$$(1 - B_k)V(\bar{z}_k) - Cov(\bar{y}_k(\bar{s}), \bar{z}_k) = \tau^2 - \tau^2 = 0. \tag{5.84}$$

Also, using $\sum_{l=1}^{L} c_{kl}\boldsymbol{x}_l = \mathbf{0}$, and (5.83), one gets

$$\begin{aligned} &Cov\left(\tilde{\boldsymbol{b}}, \sum_{l=1}^{L} c_{kl}\bar{z}_l\right) \\ &= \left(\sum_{l=1}^{L}(1 - B_l)\boldsymbol{x}_l\boldsymbol{x}_l^T\right)^{-1} Cov\left(\sum_{l=1}^{L}(1 - B_l)\bar{z}_l\boldsymbol{x}_l, \sum_{l=1}^{L} c_{kl}\bar{z}_l\right) \\ &= \left(\sum_{l=1}^{L}(1 - B_l)\boldsymbol{x}_l\boldsymbol{x}_l^T\right)^{-1} \sum_{l=1}^{L}(1 - B_l)V(\bar{z}_l)c_{kl}\boldsymbol{x}_l \\ &= \left(\sum_{l=1}^{L}(1 - B_l)\boldsymbol{x}_l\boldsymbol{x}_l^T\right)^{-1} \tau^2 \sum_{l=1}^{L} c_{kl}\boldsymbol{x}_l \\ &= \mathbf{0}. \end{aligned} \tag{5.85}$$

Combining (5.82), (5.84) and (5.85), one gets (5.80). This completes the proof of the lemma. □

In the situation when λ is unknown, in part (III) of the hierarchical model, one also assigns the gamma$(c/2, d/2)$ distribution to ΛR. Then, conditional on $\boldsymbol{z}$ and λ, the predictive distribution of $\boldsymbol{y}(\bar{s})$ is the same multivariate-t given in Theorem 5.4. To find the predictive distribution of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$, one needs in addition the conditional pdf of λ given $\boldsymbol{z}$. Integrating with respect to $\boldsymbol{\theta}$, r and $\boldsymbol{y}(\bar{s})$ that the conditional pdf of λ given $\boldsymbol{z}$ is

$$\begin{aligned} f(\lambda|\boldsymbol{z}) &\propto f(\boldsymbol{z},\lambda) \propto \lambda^{(L-1)/2} \\ &\times |\boldsymbol{E}|^{-1/2}|\boldsymbol{F}_{22}|^{-1/2}[Q(\boldsymbol{z}) + a + c\lambda]^{-(n_T+b+d-p)/2}. \end{aligned} \tag{5.86}$$

But, using Exercise 2.6 of Rao (1973), one gets

$$|\boldsymbol{F}_{22}| = |\boldsymbol{E} - Diag(N_1 - n_1, \ldots, N_L - n_L)||\boldsymbol{E}|^{-1}. \tag{5.87}$$

Hence, using the same exercise,

$$\begin{aligned} |\boldsymbol{E}|^{-1/2}|\boldsymbol{F}_{22}|^{-1/2} &= |Diag(\lambda + n_1, \ldots, \lambda + n_L) - \lambda \boldsymbol{P}_{\boldsymbol{X}}|^{-1/2} \\ &= |Diag(\lambda + n_1, \ldots, \lambda + n_L)|^{1/2} \\ &\times \left|\sum_{k=1}^{L}(1 - B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right|^{1/2} / |\boldsymbol{X}^T\boldsymbol{X}|^{1/2}. \end{aligned} \tag{5.88}$$

From (5.86)–(5.88), the conditional pdf of λ given $\boldsymbol{z}$ is

$$\begin{aligned} f(\lambda|\boldsymbol{z}) &\propto \lambda^{(L-p)/2}\prod_{k=1}^{L}(M + n_k)^{-1/2}\left|\sum_{k=1}^{L}(1 - B_k)\boldsymbol{x}_k\boldsymbol{x}_k^T\right|^{1/2} \\ &\times [Q(\boldsymbol{z}) + a + c\lambda]^{-(n_T+b+d-p)/2}. \end{aligned} \tag{5.89}$$

This leads to the following theorem.

Theorem 5.6 *Consider the hierarchical model given in (I)–(III) of Theorem 5.4 with a gamma$(c/2, d/2)$ distribution for λ. Then the posterior distribution of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$ and λ is the same multivariate-t as given in Theorem 5.4, while the posterior pdf of λ given $\boldsymbol{z}$ is given in (5.89).*

The posterior mean vector and the variance–covariance matrix of $\boldsymbol{\mu}$ are now obtained by using the iterated formulas for conditional expectations, conditional variances and conditional covariances. From (5.61) one gets,

$$E(\mu_k|\boldsymbol{z}) = \bar{z}_k - f_k E[B_k(\bar{z}_k - \boldsymbol{x}_k^T\tilde{\boldsymbol{b}})|u\boldsymbol{z}], \tag{5.90}$$

where the conditional expectation is based on the conditional pdf given in (5.88). Similarly, using Theorem 5.4, the properties of the multivariate-t distribution, (5.88), and the iterated formulas for conditional variances and conditional covariances, one finds $V(\mu_k|\boldsymbol{z})$. The details are left as an exercise.

We illustrate the results of this section with a small dataset. The analysis is taken from Ghosh and Lahiri (1992). The given data is a slightly enlarged version of what is given in Stroud (1987). The survey concerned students attending Queen's University at Kingston, Ontario, from 15 municipalities in Canada. The students were asked the following question : 'How many trips home do you estimate you will have taken by the end of the academic year'? The 15 municipalities are treated as 15 local areas, and the total number of students who responded as the finite population. As there is no reason to believe that the Arts and Science students are heterogeneous from the remaining students, the Arts and Science students are treated as samples. In this way, samples of different sizes are generated from the 15 municipalities. The 15 municipalities with their population and sample sizes given within parentheses are (1) Belleville (3, 3), (2) Brampton (3, 2), (3) Brockville (3, 2), (4) Calgary (5, 4), (5) London (3, 1), (6) Mississagua (4, 2), (7) Montreal (4, 3), (8) Oakville (3, 3), (9) Oshawa (3, 1), (10) Ottawa (25, 11), (11) Pembroke (3, 2), (12) Sault St. Marie (14, 3), (13) Sudburg (3, 1), (14) Toronto (31, 17) and (15) Vancouver (5, 4).

Following Stroud (1987), a simple linear regression is fitted. The auxiliary variable is taken as a function of the road distance between the municipality and Kingston. The plot of log mean number of trips versus the log distances for all the 15 municipalities exhibited a reasonably modest scatter about linearity, the least squares value of the slope being -0.494. Thus a reasonable x-variable is the $-1/2$ power of the road distance of the municipality from Kingston. First, a regression line is fitted with a slope as well as an intercept. In this case, $\boldsymbol{x}_k^T = (1, x_k)$, $k = 1, \ldots, 15$. Since infinite distance from Kingston implies zero trips, a second regression line is fitted without the intercept term. Thus, in the first case, $p = 2$, while for the second case $p = 1$. For R and ΛR, several gamma distributions were tried which involved using inverse J and Bell-shaped distributions or any combinations thereof. The findings led essentially to the same values of the predictive means and s.d.'s. We present below two tables one for $p = 2$ and the other for $p = 1$. In each case, $a = g = 2$, $c = 0$, and $d = 2$. The tables provide for each of

Table 5.1 *True means, sample means, x-values, predictive means (e_k), predictive s.d.'s (s_k) for the 15 municipalities when $p = 2$.*

Area (k)	True mean	Sample mean	x_k	e_k	s_k
1	12.33	12.33	0.1125	12.33	0.00
2	5.00	6.00	0.0594	5.81	1.63
3	10.67	13.50	0.1118	12.79	1.68
4	1.20	1.25	0.0171	1.12	1.00
5	4.00	3.00	0.0476	3.71	2.33
6	3.75	2.00	0.0606	4.52	1.74
7	3.00	4.00	0.0579	4.31	1.22
8	4.33	4.33	0.0587	4.33	0.00
9	5.33	6.00	0.0712	5.17	2.32
10	6.00	7.64	0.0774	7.54	0.84
11	3.00	3.50	0.0634	4.29	1.63
12	3.00	3.67	0.0337	3.37	1.23
13	2.67	4.00	0.0413	3.57	2.34
14	5.90	5.59	0.0624	5.66	0.67
15	1.60	1.25	0.0149	1.07	1.00

the 15 municipalities the true population means, the sample averages, the x-values, the predictive means, and the predictive s.d.'s. The sum of squares of the deviations between the true means and the sample means is 19.8966, while the sum of squares of the deviations between the true means and the HB estimates are given respectively by 12.8967 for $p = 2$, and 12.9144 for $p = 1$. Also, the HB estimates for the individual municipalities are quite similar for $p = 2$ and $p = 1$.

5.5 Auxiliary information II

5.5.1 A special model with two variance components

In this section, we consider the hierarchical Bayes (HB) analogue of the results considered in Section 4.6 of the previous chapter. We start with the basic model given in (4.168). However, in addition, independent priors are assigned to $\boldsymbol{B}$, $R = \sigma^{-2}$ and $W = \sigma_v^{-2}$ as follows:

The distributions of $\boldsymbol{B}$, R, and W are independent with $\boldsymbol{B} \sim$

Table 5.2 *True means, sample means, x-values, predictive means (e_k), predictive s.d.'s (s_k) for the 15 municipalities when $p = 1$.*

Area (k)	True mean	Sample mean	x_k	e_k	s_k
1	12.33	12.33	0.1125	12.33	0.00
2	5.00	6.00	0.0594	5.88	1.62
3	10.67	13.50	0.1118	12.53	1.65
4	1.20	1.25	0.0171	1.32	0.97
5	4.00	3.00	0.0476	4.01	2.29
6	3.75	2.00	0.0606	4.62	1.73
7	3.00	4.00	0.0579	4.37	1.21
8	4.33	4.33	0.0587	4.33	0.00
9	5.33	6.00	0.0712	5.16	2.32
10	6.00	7.64	0 .0774	7.47	0.83
11	3.00	3.50	0 .0634	4.34	1.62
12	3.00	3.67	0 .0337	3.55	1.21
13	2.67	4.00	0 .0413	3.94	2.29
14	5.90	5.59	0 .0624	5.73	0.64
15	1.60	1.25	0 .0149	1.28	0.96

uniform(R^p), $R \sim$ gamma($a_0/2, g_0/2$) and $W \sim$ gamma($a_1/2$, $g_1/2$).

Once again, our target is to find the conditional distribution of the unsampled observations $y(\tilde{s})$ given the observed values $\boldsymbol{z}$. With this end, first write $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_L^T)^T$. The joint (improper) pdf of $\boldsymbol{y}$, $\boldsymbol{B}$, R and W is given by

$$
\begin{aligned}
f(\boldsymbol{y}, \boldsymbol{b}, r, w) \propto & \prod_{k=1}^{L} |r^{-1}\boldsymbol{D}_k + w^{-1}c_k\boldsymbol{J}_{n_k}|^{-1/2} \\
& \times \exp\Bigg[-\sum_{k=1}^{L} (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{b})^T (r^{-1}\boldsymbol{D}_k + w^{-1}c_k\boldsymbol{J}_{n_k})^{-1} \\
& \times (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{b})/2 \Bigg] \\
& \times exp(-a_0 r/2) r^{g_0/2-1} exp(-a_1 w/2) w^{g_1/2-1}.
\end{aligned}
\tag{5.91}
$$

We now reparametrize $W = \Lambda R$. Then the joint pdf of $\boldsymbol{y}$, $\boldsymbol{B}$, R and Λ is given by

$$\begin{aligned} f(\boldsymbol{y},\boldsymbol{b},r,\lambda) &\propto r^{N_T/2}\prod_{k=1}^{L}|\boldsymbol{D}_k+\lambda^{-1}c_k\boldsymbol{J}_{n_k}|^{-1/2} \\ &\times \exp\Bigg[-r\sum_{k=1}^{L}(\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{b})^T(\boldsymbol{D}_k+\lambda^{-1}c_k\boldsymbol{J}_{n_k})^{-1} \\ &\times (\boldsymbol{y}_k-\boldsymbol{X}_k\boldsymbol{b})/2\Bigg] \\ &\times exp[-r(a_0+a_1\lambda)/2]r^{(g_0+g_1)/2-1}\lambda^{g_1/2-1} \\ &= r^{N_T/2}\prod_{k=1}^{L}|\boldsymbol{D}_k+\lambda^{-1}c_k\boldsymbol{J}_{n_k}|^{-1/2} \\ &\times exp[-r(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})/2] \\ &\times exp[-r(a_0+a_1\lambda)/2]r^{(g_0+g_1)/2-1}\lambda^{g_1/2-1}, \quad (5.92) \end{aligned}$$

where $\Sigma = \bigoplus_{k=1}^{L}(\boldsymbol{D}_k+\lambda^{-1}c_k\boldsymbol{J}_{n_k})^{-1}$. Now calculate,

$$\begin{aligned} &(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})^T\Sigma^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) \\ &= (\boldsymbol{b}-(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{y})^T(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X}) \\ &\times (\boldsymbol{b}-(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{y})+\boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{y}, \quad (5.93) \end{aligned}$$

where

$$\boldsymbol{Q} = \Sigma^{-1}-\Sigma^{-1}\boldsymbol{X}(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Sigma^{-1}. \quad (5.94)$$

Now using (5.92) and (5.93), the joint pdf of $\boldsymbol{y}$, R and Λ is given by

$$\begin{aligned} f(\boldsymbol{y},r,\lambda) &\propto |\Sigma|^{-1/2}|\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X}|^{-1/2}r^{N_T+g_0+g_1/2-1} \\ &\times exp[-r(a_0+a_1\lambda+\boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{y})/2]\lambda^{g_1/2-1}. \quad (5.95) \end{aligned}$$

Now integrating with respect to r, the joint pdf of $\boldsymbol{y}$, and Λ is given by

$$\begin{aligned} f(\boldsymbol{y},\lambda) &\propto |\Sigma|^{-1/2}|\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X}|^{-1/2} \\ &\times (a_0+a_1\lambda+\boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{y})^{-(N_T+g_0+g_1-1)/2}\lambda^{g_1/2-1}. \quad (5.96) \end{aligned}$$

Next, partition $\boldsymbol{y}$ into $\boldsymbol{y}^T = (\boldsymbol{z}^T, \boldsymbol{y}^T(\bar{s}))$, where, as before, $\boldsymbol{z}$ denotes the vector of characteristics for the observed units, while $\boldsymbol{y}(\bar{s})$ denotes the vector of characteristics for the unobserved units. Similarly, partition $\boldsymbol{X}$ into $\boldsymbol{X}^T = (\boldsymbol{X}(s)^T, \boldsymbol{X}(\bar{s})^T)$, and partition

Σ as

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Now, using a standard formula for partitioned matrices (e.g. Searle, 1971, p. 46), one gets

$$y^T\Sigma^{-1}y = z^T\Sigma_{11}^{-1}z + (y(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}z)^T\Sigma_{22.1}^{-1}(y(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}z), \tag{5.97}$$

where $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. Similarly,

$$\begin{aligned} y^T\Sigma^{-1}X &= z^T\Sigma_{11}^{-1}X(s) + (y(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}z)^T\Sigma_{22.1}^{-1} \\ &\times (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s)) \\ &= t_1^T + t_2^T \text{ (say);} \end{aligned} \tag{5.98}$$

$$\begin{aligned} X^T\Sigma^{-1}X &= X^T(s)\Sigma_{11}^{-1}X(s) + (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s))^T \\ &\times \Sigma_{22.1}^{-1}(X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s)). \end{aligned} \tag{5.99}$$

Using a standard matrix inversion formula (see e.g. Exercise 2.9, p. 33 of Rao (1973)), one gets

$$\begin{aligned} (X^T\Sigma^{-1}X)^{-1} &= (X^T(s)\Sigma_{11}^{-1}X(s))^{-1} \\ &- (X^T(s)\Sigma_{11}^{-1}X(s))^{-1} \\ &\times (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s))^T \\ &\times \{\Sigma_{22.1} + (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s)) \\ &\times (X^T(s)\Sigma_{11}^{-1}X(s))^{-1} \\ &\times (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s))^T\}^{-1} \\ &\times (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s))(X^T(s)\Sigma_{11}^{-1}X(s))^{-1} \\ &= (X^T(s)\Sigma_{11}^{-1}X(s))^{-1} - (X^T(s)\Sigma_{11}^{-1}X(s))^{-1} \\ &\times (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s))^T G^{-1} \\ &\times (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s))(X^T(s)\Sigma_{11}^{-1}X(s))^{-1} \\ &= M_1 - M_2 \text{ (say),} \end{aligned} \tag{5.100}$$

where

$$\begin{aligned} G &= \Sigma_{21} + (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s))(X^T(s)\Sigma_{11}^{-1}X(s))^{-1} \\ &\times (X(\bar{s}) - \Sigma_{21}\Sigma_{11}^{-1}X(s))^T. \end{aligned} \tag{5.101}$$

Now, after some simplification,

$$\begin{aligned} &y^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y \\ &= t_1^T(M_1 - M_2)t_1 + t_2^T(M_1 - M_2)t_2 + 2t_1^T(M_1 - M_2)t_2. \end{aligned}$$

(5.102)

Then, writing

$$\boldsymbol{K} = \boldsymbol{\Sigma}_{11}^{-1} - \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}(s)(\boldsymbol{X}^T(s)\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}(s))^{-1}\boldsymbol{X}^T(s)\boldsymbol{\Sigma}_{11}^{-1}, \quad (5.103)$$

one gets

$$\boldsymbol{t}_1^T\boldsymbol{M}_1\boldsymbol{t}_1 = \boldsymbol{z}^T(\boldsymbol{\Sigma}_{11}^{-1} - \boldsymbol{K})\boldsymbol{z}. \quad (5.104)$$

Next, writing

$$\boldsymbol{M} = \boldsymbol{\Sigma}_{21}\boldsymbol{K} + \boldsymbol{X}(\bar{s})(\boldsymbol{X}^T(s)\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}(s))^{-1}\boldsymbol{X}^T(s)\boldsymbol{\Sigma}_{11}^{-1} \quad (5.105)$$

one gets

$$\boldsymbol{t}_1^T\boldsymbol{M}_2\boldsymbol{t}_1 = (\boldsymbol{M}\boldsymbol{z} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z})^T\boldsymbol{G}^{-1}(\boldsymbol{M}\boldsymbol{z} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z}); \quad (5.106)$$

$$\boldsymbol{t}_2^T\boldsymbol{M}_1\boldsymbol{t}_2 = (\boldsymbol{y}(\bar{s}) - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z})^T(\boldsymbol{\Sigma}_{22.1}^{-1}\boldsymbol{G}\boldsymbol{\Sigma}_{22.1}^{-1} - \boldsymbol{\Sigma}_{22.1}^{-1}) \times (\boldsymbol{y}(\bar{s}) - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z}); \quad (5.107)$$

$$\boldsymbol{t}_2^T\boldsymbol{M}_2\boldsymbol{t}_2 = (\boldsymbol{y}(\bar{s}) - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z})^T(\boldsymbol{\Sigma}_{22.1}^{-1}\boldsymbol{G}\boldsymbol{\Sigma}_{22.1}^{-1} - 2\boldsymbol{\Sigma}_{22.1}^{-1} + \boldsymbol{G}^{-1})(\boldsymbol{y}(\bar{s}) - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z}); \quad (5.108)$$

$$\boldsymbol{t}_1^T\boldsymbol{M}_1\boldsymbol{t}_2 = (\boldsymbol{M}\boldsymbol{z} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z})^T\boldsymbol{\Sigma}_{22.1}^{-1}(\boldsymbol{y}(\bar{s}) - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z}); \quad (5.109)$$

$$\boldsymbol{t}_1^T\boldsymbol{M}_2\boldsymbol{t}_2 = (\boldsymbol{M}\boldsymbol{z} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z})^T(\boldsymbol{\Sigma}_{22.1}^{-1} - \boldsymbol{G}^{-1})(\boldsymbol{y}(\bar{s}) - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{z}). \quad (5.110)$$

Next, after much algebraic manipulation, one gets

$$\boldsymbol{y}^T\boldsymbol{Q}\boldsymbol{y} = \boldsymbol{z}^T\boldsymbol{K}\boldsymbol{z} + (\boldsymbol{y}(\bar{s}) - \boldsymbol{M}\boldsymbol{z})^T\boldsymbol{G}^{-1}(\boldsymbol{y}(\bar{s}) - \boldsymbol{M}\boldsymbol{z}). \quad (5.111)$$

Recall $n_T = \sum_{k=1}^{L} n_k$. Based on our calculations in (5.91)–(5.111), we have now proved the main thorem of this section as follows:

Theorem 5.7 *Assume $n_T + g_0 + g_1 > p + 2$. Then, conditional on $\Lambda = \lambda$, and $\boldsymbol{z}$, $\boldsymbol{y}(\bar{s})$ is distributed as multivariate-t with degrees of freedom $n_T + g_0 + g_1 - p$, location-vector $\boldsymbol{M}\boldsymbol{z}$, and scale-matrix $(n_T + g_0 + g_1 - p)^{-1}(a_0 + a_1\lambda + \boldsymbol{z}^T\boldsymbol{K}\boldsymbol{z})\boldsymbol{G}$. Also, the conditional pdf of Λ given $\boldsymbol{z}$ is*

$$f(\lambda|\boldsymbol{z}) \propto |\Sigma_{11}|^{-1/2}|\boldsymbol{X}^T(s)\Sigma_{11}^{-1}\boldsymbol{X}(s)|^{-1/2}\lambda^{g_1/2-1} \times (a_0 + a_1\lambda + \boldsymbol{z}^T\boldsymbol{K}\boldsymbol{z})^{-(n_T+g_0+g_1-p)/2}\lambda^{g_1/2-1} \quad (5.112)$$

Based on the above theorem, the posterior mean and the posterior

variance of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$ are given respectively by

$$E[y(\bar{s})|\boldsymbol{z}] = E(\boldsymbol{M}|\boldsymbol{z})\boldsymbol{z}; \tag{5.113}$$

$$\begin{aligned} V[(y(\bar{s})|\boldsymbol{z}] &= V[\boldsymbol{M}\boldsymbol{z}|\boldsymbol{z}] \\ &+ (n_T + g_0 + g_1 - p)^{-1} \\ &\times E[\{a_0 + a_1\lambda + \boldsymbol{z}^T\boldsymbol{K}\boldsymbol{z}\}\boldsymbol{G}|\boldsymbol{z}]. \end{aligned} \tag{5.114}$$

The posterior means and variances of $\bar{y}_k(\bar{s})$, $k = 1, \ldots, L$ are now obtained as

$$E[\bar{y}_k(\bar{s})|\boldsymbol{z}] = c_k^T E[y(\bar{s})|\boldsymbol{z}]; \tag{5.115}$$

$$V[\bar{y}_k(\bar{s})|\boldsymbol{z}] = c_k^T V[(y(\bar{s})|\boldsymbol{z}]c_k, \tag{5.116}$$

where $c_k^T = (\boldsymbol{0}^T, \ldots, \boldsymbol{0}^T, \boldsymbol{1}_{N_k - n_k}^T, \boldsymbol{0}^T, \ldots, \boldsymbol{0}^T)$. The posterior means and variances of μ_k's are then found using (5.115) and (5.116).

Using the above results, we next analyse a dataset where the objective is to predict areas under corn and soybeans for 12 counties in North Central Iowa based on the 1978 June Enumerative Survey as well as LANDSAT satellite data. The dataset appears in Battese *et al.* (1988), who conducted a variance components analysis for this problem. The background for this problem is as follows.

The USDA Statistical Reporting Service field staff determined the area of corn and soybeans in 37 sample segments (each segment was about 250 hectares) of 12 counties in North Central Iowa by interviewing farm operators. Based on LANDSAT readings obtained during August and September 1978, USDA procedures were used to classify the crop cover for all pixels (a term for a picture element of about 0.45 hectares) in the 12 counties. The number of segments in each county, the number of hectares of corn and soybeans (as reported in the June Enumerative Survey), the number of pixels classified as corn and soybeans for each sample segment, and the county mean number of pixels classified as corn and soybeans (the total number of pixels classified as that crop divided by the number of segments in that county) are reported in Table 1 of Battese *et al.* (1988). In order to make our results comparable to theirs, the second segment in Hardin County was ignored.

We shall be concerned with estimation of soybeans for the 12 counties. Battese *et al.* (1988) considered the model

$$y_i = b_0 + b_1 x_{ki1} + b_2 x_{ki2} + v_k + e_i, \tag{5.117}$$

for a unit i belonging to stratum k. Here y_i is the reported num-

ber of hectares of soybeans for unit i, x_{ki1} (x_{ki2}) is the number of pixels classified as corn (soybeans) for the ith unit which belongs to stratum k. They assumed $E(v_k) = E(e_i) = 0$, $V(v_k) = (\lambda r)^{-1}$, $V(e_i) = r^{-1}$, $Cov(v_k, e_i) = 0$, $Cov(v_k, v_{k'}) = 0$ for $k \neq k'$, $Cov(e_i, e_{i'}) = 0$ for $i \neq i'$. Such a model is usually referred to as a **nested error regression model.**

The method of Battese *et al.* is now briefly described. First, assuming λ and r to be known, obtain BLUP's of $b_0 + b_1\bar{x}_{k1p} + b_2\bar{x}_{k2p}$, where $\bar{x}_{kjp} = N_k^{-1}\sum_{i \in str_k} x_{kij}$; $j = 1, 2$. Then, using Henderson's Method III, obtain estimates of the variance components, and the final predictors will involve the estimated variance components. Henderson's method being an ANOVA method can lead to negative estimates of the variance components. If this were the case, set it equal to zero.

For applying the Bayesian method described earlier in this section, first observe that

$$\boldsymbol{\Sigma}_{11} = Diag(\boldsymbol{I}_{n_1} + \lambda^{-1}\boldsymbol{J}_{n_1}, \ldots, \boldsymbol{I}_{n_L} + \lambda^{-1}\boldsymbol{J}_{n_L})$$

so that $|\boldsymbol{\Sigma}_{11}| = \prod_{i=1}^{L}\{(\lambda + n_i)/\lambda\}$. Also, writing

$$\bar{\boldsymbol{x}}_k = n_k^{-1} \sum_{i \in str_k, i \in s} \boldsymbol{x}_{ki}, \; k = 1, \ldots, L,$$

one gets

$$\begin{aligned} \boldsymbol{X}(s)^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}(s) &= \sum_{k=1}^{L}\sum_{i \in str_k, i \in s} \boldsymbol{x}_{ki}\boldsymbol{x}_{ki}^T \\ &- \sum_{k=1}^{L} n_k^2(n_k + \lambda)^{-1}\bar{\boldsymbol{x}}_k\bar{\boldsymbol{x}}_k^T = \boldsymbol{H}(\lambda), \text{ (say).} \end{aligned} \tag{5.118}$$

Next calculate

$$\begin{aligned} \boldsymbol{z}^T\boldsymbol{K}\boldsymbol{z} &= \sum_{k=1}^{L}\sum_{i \in str_k, i \in s}(z_i - \bar{z}_k)^2 + \lambda\sum_{k=1}^{L} n_k(n_k + \lambda)^{-1}\bar{z}_k^2 \\ &- \left\{\sum_{k=1}^{L}\sum_{i \in str_k, i \in s} \boldsymbol{x}_{ki}(z_i - n_k(n_k + \lambda)^{-1}\bar{z}_k)\right\}^T \boldsymbol{H}^{-1}(\lambda) \end{aligned}$$

$$\times \left\{ \sum_{k=1}^{L} \sum_{i \in str_k, i \in s} x_{ki}(z_i - n_k(n_k + \lambda)^{-1}\bar{z}_k) \right\}$$
$$= Q_0(\lambda), \text{ (say)}. \tag{5.119}$$

The conditional pdf $f(\lambda|\boldsymbol{z})$ given in (5.112) simplifies to

$$f(\lambda|\boldsymbol{z}) \propto \lambda^{(L+g_1)/2-1} \prod_{k=1}^{L} (n_k + \lambda)^{-1/2} |\boldsymbol{H}(\lambda)|^{-1/2}$$
$$\times (a_0 + a_1\lambda + Q_0(\lambda))^{-(n_T+g_0+g_1-p)/2}. \tag{5.120}$$

The posterior means and variances of the μ_k's are now obtained from (5.115) and (5.116), respectively, using (5.118)-(5.120).

Remark 1. Let $V_1(\boldsymbol{z})$ and $V_2(\boldsymbol{z})$ denote respectively the variance of the conditional expectation, and expectation of the conditional variance of the finite population means. A naive empirical Bayes approach effectively ignores V_1 and can lead to serious underestimates of the variance. An HB procedure, on the other hand, rectifies this deficiency. Battese *et al.* have a frequentist approach which also incorporates the uncertainty of estimating the variance components into account.

We provide below the HB predictors (e_{HB}), the EB predictors (e_{EB}), and also the predictors obtained by Battese *et al.* (e_{BHF}), and the associated standard errors s_{HB}, s_{EB} and s_{BHF} respectively for mean areas under soybeans in the 12 counties. For finding the HB predictors, we have used the improper priors with $a_0 = a_1 = 0.005$ and $g_0 = g_1 = 0$. Note that the EB predictors are obtained by replacing λ with its estimate obtained by Henderson's Method III in $E[\mu_k|\boldsymbol{z}, \lambda]$.

As one might anticipate, e_{HB}, e_{EB} and e_{BHF} are all extremely close as point predictors; e_{BHF} differs from e_{EB} because it uses a different estimate of λ. Also, the naive EB estimator, in general, underestimates the standard error in comparison with the HB estimator. With the exception of Hamilton County, s_{EB} is always smaller than or equal to s_{HB}. The difference can be significant as evidenced from the figures given in Humboldt County where s_{EB} is about 10% smaller than s_{HB}.

However, both s_{HB} and s_{BHF} are very good estimates of standard errors. In this example, while s_{BHF} is never smaller than s_{HB} by more than 6.1%, it can exceed s_{HB} by about 9.7%.

Table 5.3 *The predicted hectares of soybeans and standard errors.*

County	e_{HB}	e_{EB}	e_{BHF}	s_{HB}	s_{EB}	s_{BHF}
Cerro Gordo	78.8	78.2	77.5	11.7	11.6	12.7
Franklin	67.1	65.9	64.8	8.2	7.5	7.8
Hamilton	94.4	94.6	95.0	11.2	11.4	12.4
Hancock	100.4	100.8	101.1	6.2	6.1	6.3
Hardin	75.4	75.1	74.9	6.5	6.4	6.6
Humboldt	81.9	80.6	79.2	10.4	9.3	10.0
Kossuth	118.2	119.2	120.2	6.6	6.0	6.2
Pocahontas	113.9	113.7	113.8	7.5	7.5	7.9
Webster	110.0	109.7	109.6	6.6	6.6	6.8
Winnebago	97.3	98.0	98.7	7.7	7.5	7.9
Worth	87.8	87.2	86.6	11.1	11.1	12.1
Wright	111.9	112.4	112.9	7.7	7.6	8.0

5.5.2 *The general mixed linear model with several variance components*

A more general hierarchical model was considered in Datta and Ghosh (1991). They considered the model

(A) conditional on $\boldsymbol{b} = (b_1, \ldots, b_p)^T$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_t)^T$ and r, let

$$\boldsymbol{y} \sim N(\boldsymbol{Xb}, r^{-1}(\boldsymbol{\Psi} + \boldsymbol{WD}(\boldsymbol{\lambda})\boldsymbol{W}^T)),$$

where $\boldsymbol{y}$ is $N_T \times 1$;

(B) $\boldsymbol{B}$, R and $\boldsymbol{\lambda}$ have a certain prior distribution proper or improper.

Stage (A) of this model can be identified as a general mixed linear model. To see this, write

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Wv} + \boldsymbol{e}, \tag{5.121}$$

where $\boldsymbol{e}$ and $\boldsymbol{v}$ are mutually independent with $\boldsymbol{e} \sim N(\boldsymbol{0}, r^{-1}\boldsymbol{\Psi})$ and $\boldsymbol{v} \sim N(\boldsymbol{0}, r^{-1}\boldsymbol{D}(\boldsymbol{\lambda}))$, where $\boldsymbol{e}$ is $N_T \times 1$, and $\boldsymbol{v}$ is $q \times 1$; also, $\boldsymbol{X}(N_T \times p)$ and $\boldsymbol{W}(N_T \times q)$ are known design matrices; $\boldsymbol{\Psi}$ is a p.d. matrix, and $\boldsymbol{D}(\boldsymbol{\lambda})$ $(q \times q)$ is a p.d. matrix which is structurally known except possibly for some unknown $\boldsymbol{\lambda}$. In the examples which follow, $\boldsymbol{\lambda}$ involves the ratios of the variance components. Sometimes we will denote $\boldsymbol{D}(\boldsymbol{\lambda})$ by $\boldsymbol{D}$ when $\boldsymbol{\lambda}$ is known.

Once again, we partition $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{W}$ and $\boldsymbol{e}$ as

$$\begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{y}(\bar{s}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}(s) \\ \boldsymbol{X}(\bar{s}) \end{pmatrix} \boldsymbol{b} + \begin{pmatrix} \boldsymbol{W}(s) \\ \boldsymbol{W}(\bar{s}) \end{pmatrix} \boldsymbol{v} + \begin{pmatrix} \boldsymbol{e}(s) \\ \boldsymbol{e}(\bar{s}) \end{pmatrix}, \quad (5.122)$$

where s and $\bar{s}$ refer as before to the sampled and unsampled elements respectively. For simplicity, it is assumed that $\text{rank}(\boldsymbol{X}(s)) = p$.

The objective is once again to find the predictive distribution of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$. However, prior to that, we identify some of the existing models for small area estimation as special cases of (5.122). In addition to the usual notations like $\mathbf{1}_u$, $\boldsymbol{J}_u$ and I_u, we need the additional notations $col_{1\le i\le L}(\boldsymbol{B}_i) = (\boldsymbol{B}_1^T, \ldots, \boldsymbol{B}_L^T)^T$ and

$$\bigoplus_{i=1}^{L} \boldsymbol{A}_i = \begin{bmatrix} \boldsymbol{A}_1 \ldots \boldsymbol{0} \\ \boldsymbol{0} \ldots \boldsymbol{A}_L \end{bmatrix}.$$

First, the nested error regression model considered earlier in this section is easily seen to be a special case of this general model with $\boldsymbol{W}(s) = \bigoplus_{i=1}^{L} \mathbf{1}_{n_i}$, $\boldsymbol{W}(\bar{s}) = \bigoplus_{i=1}^{L} \mathbf{1}_{N_i - n_i}$, $\boldsymbol{\Psi} = \boldsymbol{I}_{N_T}$, $t = 1$, $\boldsymbol{\lambda} = \lambda$ and $\boldsymbol{D}(\boldsymbol{\lambda}) = \lambda^{-1}\boldsymbol{I}_L$. In a more special case, Ghosh and Lahiri (1992) considered the model when $\boldsymbol{x}_{ki} = \boldsymbol{x}_k$ for all units i belonging to stratum k.

The random regression coefficients model of Dempster et. al. (1981) is also a special case of the model given in (5.121) or (5.122). In this setup, $\boldsymbol{X}(s)$, $\boldsymbol{X}(\bar{s})$, $\boldsymbol{\Psi}$ and $\boldsymbol{D}(\boldsymbol{\lambda})$ are the same as in the nested error regression model, but $\boldsymbol{W}(s) = \bigoplus_{i=1}^{L} [col_{i\in s} \boldsymbol{x}_{ki}^T]$ and $\boldsymbol{W}(\bar{s}) = \bigoplus_{i=1}^{L} [col_{i\in \bar{s}} \boldsymbol{x}_{ki}^T]$. Prasad and Rao (1990) considered both the nested error and the random regression coefficients models in a frequentist framework.

Next we identify a two-stage sampling model with covariates and with L strata as also a special case of the model given in (5.121) or (5.122). Suppose that the kth stratum contains M_k primary units. Suppose also that the jth primary unit within the kth stratum contains N_{kj} subunits. From the kth stratum, a sample of m_k primary units are sampled. For the j selected primary unit within the kth stratum, a sample of n_{kj} subunits are selected. For the ith unit in the kth stratum and the jth primary unit, the characteristic of interest y_i is modelled by

$$y_i = \boldsymbol{x}_{kj}^T \boldsymbol{b} + \xi_k + \eta_{kj} + e_i, \quad (5.123)$$

where the ξ_k, η_{kj} and e_i are mutually independent with the ξ_k being iid normal with mean 0 and variance $(\lambda_1 r)^{-1}$), the η_{kj} being iid normal with mean 0 and variance $(\lambda_2 r)^{-1}$) and the e_i being iid normal with mean 0 and variance r^{-1}. Here $t = 2$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$, $\boldsymbol{\Psi} = \boldsymbol{I}_{N_T}$, $\boldsymbol{D}(\boldsymbol{\lambda}) = Diag(\lambda_1^{-1}\boldsymbol{I}_L, \lambda_2^{-1}\boldsymbol{I}_M)$, where $M = \sum_{k=1}^{L} M_k$. These ideas can be directly extended to multistage sampling. We do not pursue this to avoid some cumbersome notations.

Next in this subsection, we provide the conditional pdf of $\boldsymbol{y}^T(\bar{s})$ given $\boldsymbol{z}$. We assume condition (A) given at the beginning of this subsection. In stage (B) of this model, it is assumed that

$$\boldsymbol{B},\ R,\ \Lambda_1 R, \ldots, \Lambda_t R \text{ are independently distributed}$$

with $\boldsymbol{B} \sim uniform(R^p)$, $R \sim gamma(a_0/2, g_0/2)$, $a_0 \geq 0$, $g_0 \geq 0$, $\Lambda_i R \sim gamma(a_i/2, g_i/2)$, $i = 1, \ldots, t$, with $a_i > 0$, $g_i \geq 0$ for all $i = 1, \ldots, t$.

Recall the notations $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{ij}$, $i, j = 1, 2$, $\boldsymbol{K}$, $\boldsymbol{M}$ and $\boldsymbol{G}$ of the previous section. The following theorem generalizes the main result of the previous section. The proof is analogous, and is omitted.

Theorem 5.8 *Consider the model given in (5.121) or (5.122). Assume that* $n_T + \sum_{i=0}^{t} g_i - p > 2$. *Then, conditional on* $\boldsymbol{\Lambda} = \boldsymbol{\lambda}$ *and* $\boldsymbol{Z} = \boldsymbol{z}$, $\boldsymbol{y}(\bar{s})$ *is distributed as multivariate-t with degrees of freedom* $n_T + \sum_{i=0}^{t} g_i - p$, *location parameter* $\boldsymbol{Mz}$, *and scale parameter*

$$\left(n_T + \sum_{i=0}^{t} g_i - p\right)^{-1} \left[a_0 + \sum_{i=1}^{t} a_i \lambda_i + \boldsymbol{z}^T \boldsymbol{K} \boldsymbol{z}\right] \boldsymbol{G}.$$

Also, the conditional distribution of $\boldsymbol{\Lambda}$ *given* $\boldsymbol{Z} = \boldsymbol{z}$ *has pdf*

$$\begin{aligned} f(\boldsymbol{\lambda}|\boldsymbol{z}) \quad &\propto \quad |\boldsymbol{\Sigma}_{11}|^{-1/2} |\boldsymbol{X}^T(s) \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{X}(s)|^{-1/2} \prod_{i=1}^{t} \lambda_i^{g_i/2-1} \\ &\times \quad \left[a_0 + \sum_{i=1}^{t} a_i \lambda_i + \boldsymbol{z}^T \boldsymbol{K} \boldsymbol{z}\right]^{-(n_T + \sum_{i=0}^{t} g_i - p)/2} \end{aligned} \tag{5.124}$$

Using the moments of a multivariate-t distribution, it follows now that if $n_T + \sum_{i=0}^{t} g_i - p > 2$, then

$$E[\boldsymbol{y}(\bar{s})|\boldsymbol{z}] = E(\boldsymbol{M}|\boldsymbol{z})\boldsymbol{z}; \tag{5.125}$$

$$V[\boldsymbol{y}(\bar{s})|\boldsymbol{z}] = V(\boldsymbol{M}\boldsymbol{z}|\boldsymbol{z}) + \left(n_T + \sum_{i=0}^{t} g_i - p - 2\right)^{-1} \times E\left[\left\{a_0 + \sum_{i=1}^{t} a_i\lambda_i + \boldsymbol{z}^T\boldsymbol{K}\boldsymbol{z}\right\}\boldsymbol{G}|\boldsymbol{z}\right]. \tag{5.126}$$

Using (5.125) and (5.126), it is possible to find the posterior means and variances of $\boldsymbol{\xi}(\boldsymbol{z}, \boldsymbol{y}(\bar{s})) = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{C}\boldsymbol{y}(\bar{s})$ where $\boldsymbol{A}$ and $\boldsymbol{C}$ are known matrices. The Bayes estimator of $\boldsymbol{\xi}(\boldsymbol{z}, \boldsymbol{y}(\bar{s}))$ under any quadratic loss is its posterior mean, and is given by

$$E[\boldsymbol{\xi}(\boldsymbol{z}, \boldsymbol{y}(\bar{s}))|\boldsymbol{z}] = [\boldsymbol{A} + \boldsymbol{C}E(\boldsymbol{M}|\boldsymbol{z})]\boldsymbol{z}. \tag{5.127}$$

Similarly, using (5.126), one finds

$$V[\boldsymbol{\xi}(\boldsymbol{z}, \boldsymbol{y}(\bar{s}))|\boldsymbol{z}] = \boldsymbol{C}V[\boldsymbol{y}(s')|\boldsymbol{z}]\boldsymbol{C}^T. \tag{5.128}$$

Note that when $\boldsymbol{A} = \bigoplus_{i=1}^{L} \mathbf{1}_{n_k}^T$ and $\boldsymbol{C} = \bigoplus_{i=1}^{L} \mathbf{1}_{N_k - n_k}^T$, then $\boldsymbol{\xi}(\boldsymbol{z}, \boldsymbol{y}(\bar{s}))$ reduces to the vector of population totals for the L local areas. In practice, inference related to these totals is of prime interest.

We now analyse the same radiation therapy dataset as considered in Section 4.5 using the results of Theorem 5.8. In addition to the EB and other model- and design-based estimators, we have the HB estimators as well in this section. The sampling methodology is discussed in detail in Section 4.5.

As before, let y_i denote the score for the ith patient. Although the y_i lie between 0 and 1, these are weighted averages of independent Bernoulli variables, and a normal approximation due to the CLT is not totally out of the way.

We assume the model given in (5.121) or (5.122) with $\boldsymbol{b} = \mu$, the general effect, and $\boldsymbol{x}_{ki} = 1$. As described earlier, from the kth stratum, a sample of m_k primary units is taken, while for the jth selected primary unit within the kth stratum, a sample of n_{kj} ($< N_{kj}$) subunits are selected. Once again, let $\boldsymbol{z}$ denote the vector of sample observations, $\bar{z}_{kj} = n_{kj}^{-1} \sum_{i \in str_k, i \in psu\ j \text{ in } str_k, i \in s} z_i$. Moreover, let

$$B_{kj} = \lambda_2/(\lambda_2 + n_{kj});$$

$$\bar{z}_k = \sum_{j=1}^{m_k}(1 - B_{kj})\bar{z}_{kj} / \sum_{j=1}^{m_k}(1 - B_{kj});$$

$$\alpha_k = \lambda_1 / \left(\lambda_1 + \lambda_2 \sum_{j=1}^{m_k} (1 - B_{kj}) \right);$$

$$\bar{z} = \sum_{k=1}^{5} (1 - \alpha_k) \bar{z}_k / \sum_{k=1}^{5} (1 - \alpha_k);$$

$$f_{kj} = (N_{kj} - n_{kj}) / N_{kj}.$$

Then the HB predictor of $\sum_{j=1}^{m_k} \sum_{i \in str_k, i \in psu\ j \text{ in } str_k} y_i / \sum_{j=1}^{m_k} N_{kj}$ is given by

$$\begin{aligned} e_{HB}^k &= \left(\sum_{j=1}^{m_k} N_{kj} \right)^{-1} E \left[\sum_{j=1}^{m_k} N_{kj} (1 - f_{kj} B_{kj}) \bar{z}_{kj} \right. \\ &+ \left\{ \left(\sum_{j=m_k+1}^{M_k} N_{kj} \right) + \sum_{j=1}^{m_k} N_{kj} f_{kj} B_{kj} \right\} \\ &\times \left. \{ (1 - \alpha_k) \bar{z}_k + \alpha_k \bar{z} \} | \boldsymbol{z} \right]. \end{aligned} \tag{5.129}$$

The posterior pdf of Λ given in (5.124) simplifies in this case to

$$\begin{aligned} f(\lambda_1, \lambda_2 | \boldsymbol{z}) \propto & \left(\prod_{k=1}^{L} \prod_{j=1}^{m_k} B_{kj}^{1/2} \right) \left(\prod_{k=1}^{L} \alpha_k^{1/2} \right) \left(\lambda_1 \sum_{k=1}^{L} (1 - \alpha_k) \right)^{-1/2} \\ & \times \left(s + a_0 + a_1 \lambda_1 + a_2 \lambda_2 + \sum_{k=1}^{L} K_{3k} \right. \\ & \left. - \left(\sum_{k=1}^{L} K_{2k} \right)^2 / \left(\sum_{k=1}^{L} K_{1k} \right) \right)^{-(n_T + g_0 + g_1 + g_2 - 1)/2}, \end{aligned} \tag{5.130}$$

where $L = 5$, $s = \sum_{k=1}^{L} \sum_{j=1}^{m_k} \sum_{i \in str_k, i \in s} (z_i - \bar{z}_{kj})^2$, $K_{1k} = \lambda_1 (1 - \alpha_k)$, $K_{2k} = \lambda_1 (1 - \alpha_k) \bar{z}_k$, $K_{3k} = \lambda_2 [\sum_{j=1}^{m_k} (1 - B_{kj}) \bar{z}_{kj}^2 - (1 - \alpha_k) \sum_{j=1}^{m_k} (1 - B_{kj}) \bar{z}_k^2]$. In finding the HB predictor, the prior with $a_0 = g_0 = g_1 = g_2 = 0$ and $a_1 = a_2 = 0.0005$ is used, and two-dimensional numerical integrations are carried out.

Table 5.4 provides the true means and the six different sets of estimates.

The average absolute biases of the HB estimates, the EB estimates, the design unbiased estimates, the ratio-type estimates, the

Table 5.4 *The true means μ_k's and their estimates.*

k	μ_k	e^k_{HB}	e^k_{EB}	e^k_U	e^k_R	e^k_0	e^k_{RO}
1	0.733	0.798	0.803	0.712	0.918	0.922	0.932
2	0.761	0.764	0.764	0.910	0.772	0.768	0.750
3	0.745	0.768	0.768	0.782	0.784	0.783	0.750
4	0.689	0.751	0.750	0.896	0.739	0.740	0.715
5	0.746	0.741	0.742	0.981	0.713	0.726	0.720

expansion estimates and Royall's estimates for the given dataset are given respectively by 0.0310, 0.0316, 0.1293, 0.0628, 0.0601 and 0.0484. Thus the HB estimates have a slight edge over EB estimates, and much greater edge over the other four estimates in terms of absolute bias. Also, the total sum of squared deviations of the HB estimates from the true means is 0.0085. The corresponding figures for e_{EB}, e_U, e_R, e_0 and e_{RO} turn out to be 0.0091, 0.1211, 0.0391 0.0400 and 0.0409 respectively. Hence, the percentage reduction in the total sum of squared deviations for the HB estimates is 6.6, 93.0, 78.3, 78.8 and 79.3 in comparison with the EB estimates, the design unbiased estimates, the ratio-type estimates, the expansion estimates and Royall's estimates respectively.

An EB point estimator is usually on par with the corresponding HB estimator. Thus the small improvement of the HB estimator over the EB estimator is not so surprising. However, the improvement of the HB estimator over the other four estimators is indeed startling. One possible explanation of this fact is that many of the other estimators are optimal under models which do not take into account variation in the primary sampling units. The HB model takes into account this extra source of variation, and thus produces more reliable estimators.

We have noticed already that the conditional pdf of $\boldsymbol{y}(\bar{s})$ given $\boldsymbol{z}$ cannot be obtained analytically due to the complicated posterior pdf of $\boldsymbol{\Lambda}$ given $\boldsymbol{z}$. The dataset of this subsection was analysed using direct numerical evaluation of two-dimensional integrals. This method does not work when the dimension of $\boldsymbol{\lambda}$ is very large. In such cases, Monte Carlo numerical integration seems to be the appropriate procedure.

A natural candidate for such purposes seems to be the importance sampling. To implement such a procedure, write $f(\boldsymbol{\lambda}|\boldsymbol{z}) = cu(\boldsymbol{\lambda}, \boldsymbol{z})$, where the norming constant c has to be numerically evaluated. Now, for any real valued function $h(\boldsymbol{\lambda})$,

$$\begin{aligned}\int_0^\infty \cdots \int_0^\infty & h(\boldsymbol{\lambda}) f(\boldsymbol{\lambda}|\boldsymbol{z}) d\boldsymbol{\lambda} \\ = & \int_0^\infty \cdots \int_0^\infty h(\boldsymbol{\lambda})\{u(\boldsymbol{\lambda}, \boldsymbol{z})/g(\boldsymbol{\lambda}|\boldsymbol{z})\} g(\boldsymbol{\lambda}|\boldsymbol{z}) d\boldsymbol{\lambda} \\ \div & \int_0^\infty \cdots \int_0^\infty \{u(\boldsymbol{\lambda}, \boldsymbol{z})/g(\boldsymbol{\lambda}|\boldsymbol{z})\} g(\boldsymbol{\lambda}|\boldsymbol{z}) d\boldsymbol{\lambda}, \qquad (5.131)\end{aligned}$$

where $g(\boldsymbol{\lambda}|\boldsymbol{z})$ is some 'standard' pdf from which random samples can be easily generated. Hence, $\int_0^\infty \cdots \int_0^\infty h(\boldsymbol{\lambda}) f(\boldsymbol{\lambda}|\boldsymbol{z}) d\boldsymbol{\lambda}$ can be approximated by

$$\sum_{i=1}^m h(\boldsymbol{\lambda}^i)\{u(\boldsymbol{\lambda}^i, \boldsymbol{z})/g(\boldsymbol{\lambda}^i|\boldsymbol{z})\} / \sum_{i=1}^m \{u(\boldsymbol{\lambda}^i, \boldsymbol{z})/g(\boldsymbol{\lambda}^i|\boldsymbol{z})\},$$

where the number of replicates m is very large, and $\boldsymbol{\lambda}^i$'s are generated from $g(\boldsymbol{\lambda}^i|\boldsymbol{z})$.

Unfortunately, finding $g(\boldsymbol{\lambda}^i|\boldsymbol{z})$ in the present context can be quite formidable. Even when $\boldsymbol{\lambda}$ is one-dimensional, $f(\boldsymbol{\lambda}|\boldsymbol{z})$ may turn out to be multimodal, and may defy any simple approximation. It is, therefore, natural to seek other Monte Carlo integration methods.

The simplest approach for this problem seems to be the use of the Gibbs sampling technique, originally introduced in Geman and Geman (1984), and more recently popularized by Gelfand and Smith (1990). The method is described below.

Gibbs sampling is a Markovian updating scheme. Given an arbitrary starting set of values, $U_1^0, \ldots, U_p^0$, one draws

$$\begin{aligned} U_1^1 &\sim [U_1|U_2^0, \ldots, U_p^0], \\ U_2^1 &\sim [U_2|U_1^1, U_3^0, \ldots, U_p^0], \ldots, \\ U_p^1 &\sim [U_p|U_1^1, \ldots, U_{p-1}^1], \end{aligned}$$

where $[.|.]$ denotes the relevant conditional distributions. Thus, each variable is visited in a natural order, and a cycle in this scheme requires p random variate generations. After k such iterations, one arrives at $(U_1^k, \ldots, U_p^k)$. As $k \to \infty$, $(U_1^k, \ldots, U_p^k)$ converges in distribution to $(U_1, \ldots, U_p)$. Gibbs sampling through

s replications of the aforementioned k-iterations generates s iid p-tuples $(U_{1j}^k, \ldots, U_{pj}^k)$, $j = 1, \ldots, s$; $U_1, \ldots, U_p$ could possibly be vectors in the above scheme.

One needs to find the relevant posteriors for implementation of the Gibbs sampler. Consider the special case when $\boldsymbol{\Psi} = \boldsymbol{I}_{N_T}$, and $\boldsymbol{D}(\boldsymbol{\lambda}) = Diag(\lambda_1^{-1}\boldsymbol{I}_{q_1}, \ldots, \lambda_t^{-1}\boldsymbol{I}_{q_t})$, where $\sum_{i=1}^t q_i = q$. Write $S_i = R\Lambda_i$, and correspondingly $s_i = r\lambda_i$, $i = 1, \ldots, t$. Assign a uniform prior for $\boldsymbol{B}$, a gamma$(a_0/2, g_0/2)$ prior for R, and the priors for the S_i are gamma$(a_i/2, g_i/2)$, where $\boldsymbol{B}$, R, $S_1, \ldots S_t$ are all independently distributed.

Write $\boldsymbol{v}^T = (\boldsymbol{v}_1^T, \ldots, \boldsymbol{v}_t^T)$, where $\boldsymbol{v}_i$ has dimension q_i. Based on the model given in (5.121) or (5.122), the joint posterior pdf of $\boldsymbol{y}(\bar{s})$, $\boldsymbol{B}$, $\boldsymbol{v}$, R, S_1, ..., S_t given $\boldsymbol{z}$ is given by

$$\begin{aligned} & f(\boldsymbol{y}(\bar{s}), \boldsymbol{b}, \boldsymbol{v}, r, s_1, \ldots, s_t | \boldsymbol{z}) \\ & \quad \propto r^{n_T/2} \exp[-r||\boldsymbol{z} - \boldsymbol{X}(s)\boldsymbol{b} - \boldsymbol{W}(s)\boldsymbol{v}||^2/2] \\ & \quad \times r^{(N_T - n_T)/2} \exp[-r||\boldsymbol{y}(s') - \boldsymbol{X}(s')\boldsymbol{b} - \boldsymbol{W}(s')\boldsymbol{v}||^2/2] \\ & \quad \times \prod_{i=1}^t \{w_i^{q_i/2} \exp(-w_i||\boldsymbol{v}_i||^2/2)\} \exp(-a_0 r/2) r^{g_0/2-1} \\ & \quad \times \prod_{i=1}^t \{exp(-a_i w_i/2) w_i^{g_i/2-1}\}. \end{aligned} \tag{5.132}$$

The required full conditional distributions are then given by

$$\begin{aligned} & \boldsymbol{B}|\boldsymbol{y}(\bar{s}), \boldsymbol{v}, r, s_1, \ldots, s_t, \boldsymbol{z} \\ & \quad \sim N[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{v}), r^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}]; \end{aligned} \tag{5.133}$$

$$\begin{aligned} & \boldsymbol{v}|\boldsymbol{y}(\bar{s}), \boldsymbol{b}, r, s_1, \ldots, s_t, \boldsymbol{z} \\ & \quad \sim N\left[\left(\boldsymbol{W}^T\boldsymbol{W} + \bigoplus_{l=1}^t r^{-1} w_l \boldsymbol{I}_{q_l}\right)^{-1} \boldsymbol{W}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}), \right. \\ & \qquad \left. r^{-1}\left(\boldsymbol{W}^T\boldsymbol{W} + \bigoplus_{l=1}^t r^{-1} w_l \boldsymbol{I}_{q_l}\right)^{-1}\right]; \end{aligned} \tag{5.134}$$

$$\begin{aligned} & R|\boldsymbol{y}(\bar{s}), \boldsymbol{b}, \boldsymbol{v}, s_1, \ldots, s_t, \boldsymbol{z} \\ & \quad \sim gamma((||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b} - \boldsymbol{W}\boldsymbol{v}||^2 + a_0)/2, (N_T + g_0)/2); \end{aligned} \tag{5.135}$$

$$S_i|\boldsymbol{y}(\bar{s}), \boldsymbol{b}, \boldsymbol{v}, r, s_j (j \neq i), \boldsymbol{z}$$

$$\sim \quad gamma((||\boldsymbol{v}_i||^2 + a_i)/2, (q_i + g_i)/2), i = 1, \ldots, t\,; \tag{5.136}$$

$$\begin{aligned} &\boldsymbol{y}(\bar{s})|\boldsymbol{b}, \boldsymbol{v}, r, s_1, \ldots, s_t, \boldsymbol{z} \\ &\quad \sim \quad N(\boldsymbol{X}(\bar{s})\boldsymbol{b} + \boldsymbol{W}(\bar{s})\boldsymbol{v}, r^{-1}\boldsymbol{I}_{N_T - n_T}). \end{aligned} \tag{5.137}$$

5.6 Generalized linear models

The HB analysis of the previous sections has mainly concentrated on numerical-valued variates. Often, however, the survey data are categorical, for which the HB or EB analysis suitable for continuous variates is not very appropriate. Recently, some work has begun on the HB or EB analysis of binary data. Dempster and Tomberlin (1980) and MacGibbon and Tomberlin (1989) obtain small area estimates of proportions via EB techniques while Malec *et al.* (1993) find the predictive distribution of a linear combination of binary random variables using an HB technique. Stroud (1991) also develops a general HB methodology for the analysis of binary data, and subsequently Stroud (1994) provides a comprehensive treatment of binary categorical survey data encompassing simple random, stratified, cluster and two-stage sampling as well as two-stage sampling within strata.

The binary models form a subclass of generalized linear models which are often used for a unified analysis of both discrete and continuous data. We offer in this section a general account of how GLM's can be used for simultaneous estimation of strata means. Hierarchical models are used where the response from units within different strata are assumed to follow a GLM structure, while the model parameters have a certain hierarchical prior. Needless to say that such procedures are very computer-intensive, and require numerical integration techniques for their implementation.

As before, suppose there are L strata. For a unit i belonging to stratum k, the response y_i is assumed to follow the model

$$f(y_i|\theta_k, \phi_k) = exp[\phi_k^{-1}(y_i\theta_k - \psi(\theta_k)) + \rho(y_i, \phi_k)]. \tag{5.138}$$

In addition, the y_i are assumed to be independently distributed.

Such a model is referred to as a generalized linear model (GLM), studied very extensively in McCullagh and Nelder (1989) using a frequentist approach. The density given in (5.138) is parametrized with respect to the canonical parameters θ_k and the scale parameters ϕ_k. In many standard GLM's including the binomial and

Poisson cases, the ϕ_k are known, which is going to be the case in the remainder of this section. Also, we may observe that if $\bar{z}_k = n_k^{-1} \sum_{i \in s, i \in str_k} y_i$, then $\bar{z}_1 \ldots, \bar{z}_L$ is minimal sufficient, and inference about the θ_k will be based on the $\bar{z}_k$.

The θ_k are modelled as

$$\theta_k = \boldsymbol{x}_k^T \boldsymbol{b} + u_k, \tag{5.139}$$

$k = 1, \ldots, L$, where the $\boldsymbol{x}_k$ are known design vectors, $\boldsymbol{b}$ $(p \times 1)$ is the vector of unknown regression coefficients, u_k are the random effects. It is assumed that the u_k are iid $N(0, \sigma_u^2)$.

We discuss below an HB analysis which assigns distributions to $\boldsymbol{b}$ and σ_u^2. We consider the prior under which marginally $\boldsymbol{b}$ and σ_u^2 are marginally independent with $\boldsymbol{b} \sim$ uniform(R^p) and $\sigma_u^{-2} \sim$ gamma$(a/2, g/2)$. Thus, we have the following hierarchical model:

(I) Conditional on $\boldsymbol{\theta}$, $\boldsymbol{b}$, and σ_u^2, y_i are mutually independent with a pdf given in (5.138).

(II) Conditional on $\boldsymbol{b}$, and σ_u^2, θ_k are independent $N(\boldsymbol{x}_k^T \boldsymbol{b}, \sigma_u^2)$.

(III) Marginally, $\boldsymbol{b}$, and σ_u^2 are mutually independent with $\boldsymbol{b} \sim$ $uniform(R^p)$ and $\sigma_u^{-2} \sim gamma(a/2, g/2)$.

This type of HB model, closely related to Zeger and Karim (1991), is appropriate in a number of situations. Breslow and Clayton (1993) refer to (I) and (II) as **generalized mixed effects linear model** in contrast to a fixed effects model. The latter did not get into any Bayesian consideration whatsoever, and used instead a **penalized likelihood** approach. It may be noted also that both Zeger and Karim (1991) and Breslow and Clayton (1993) considered a slightly more generalized version of this model than that given in (II). Instead of modelling the θ_k, they modelled $h(\theta_k)$ for some arbitrary monotone function h.

This particular hierarchical model should be contrasted to that of Albert (1988) which generalizes Leonard and Novick (1986). Albert's method when applied to the present setting will first use independent conjugate priors

$$\pi(\theta_k | m_k, \zeta) = exp[\zeta(m_k \theta_k - \psi(\theta_k)) + g(m_k, \zeta)] \tag{5.140}$$

for the θ_k. Next assume $h(m_k) = \boldsymbol{x}_k^T \boldsymbol{b}$ for some known monotone function h including the case when h is the identity function. In the next stage, assign priors (possibly diffuse) to the hyperparameters $\boldsymbol{b}$ and ζ. In a way, the Zeger–Karim and Breslow–Clayton approaches seem more natural in the present context since it models directly

some function of the means of the different strata, rather than modelling the prior parameters which are onestep removed from the natural parameters.

Two special cases of this model are of immense practical interest. One is the logistic regression model where $\theta_k = log(p_k/(1-p_k))$, p_k being the success probabilities in Bernoulli trials. Second is the log–linear model where $\theta_k = log(\lambda_k)$, λ_k being the Poisson means.

We first show that the prior given in (I)–(III) leads to proper posteriors for the θ_k under mild conditions. To this end, the following theorem is proved.

Theorem 5.9 *Consider the hierarchical model given in (I)–(III). Assume that* $L + g > p$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$ *are linearly independent and*

$$\int_{-\infty}^{\infty} exp[n_k \phi_k^{-1}(\bar{z}_k \theta - \psi(\theta)) + \rho(y, \phi_k)] d\theta < \infty \qquad (5.141)$$

for all $k = 1, \ldots, L$. *Then* $\boldsymbol{\theta}$ *has a proper posterior.*

Proof. Write $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)^T$. The joint posterior pdf of $\boldsymbol{\theta}$, $\boldsymbol{b}$ and σ_u^2 is

$$\begin{aligned} \pi(\boldsymbol{\theta}, \boldsymbol{b}, \sigma_u^2 | \boldsymbol{z}) &\propto \exp\left[\sum_{k=1}^{L} n_k \phi_k^{-1}(\bar{z}_k \theta_k - \psi(\theta_k))\right] (\sigma_u^2)^{-(L+g+1)/2} \\ &\times \exp\left[-\left\{a + \sum_{k=1}^{L} (\theta_k - \boldsymbol{x}_k^T \boldsymbol{b})^2\right\} / (2\sigma_u^2)\right]. \qquad (5.142) \end{aligned}$$

Writing $\boldsymbol{X}^T = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L)$, $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$, and integrating with respect to $\boldsymbol{b}$, one gets the joint posterior of $\boldsymbol{\theta}$ and σ_u^2 is

$$\begin{aligned} \pi(\boldsymbol{\theta}, \sigma_u^2 | \boldsymbol{z}) &\propto \exp\left[\sum_{k=1}^{L} n_k \phi_k^{-1}\{\bar{z}_k \theta_k - \psi(\theta_k)\}\right] (\sigma_u^2)^{-(L+g+1-p)/2} \\ &\times exp[-\{a + \boldsymbol{\theta}^T (\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{\theta}\} / (2\sigma_u^2)]. \qquad (5.143) \end{aligned}$$

Finally, integrating with respect to σ_u^2, the posterior of $\boldsymbol{\theta}$ is

$$\begin{aligned} \pi(\boldsymbol{\theta} | \boldsymbol{z}) &\propto \exp\left[\sum_{k=1}^{L} n_k \phi_k^{-1}(\bar{z}_k \theta_k - \psi(\theta_k))\right] \\ &\times [a + \boldsymbol{\theta}^T (\boldsymbol{I}_p - \boldsymbol{P}_{\boldsymbol{X}}) \boldsymbol{\theta}]^{-(L+g-p)/2} \\ &\leq \exp\left[\sum_{k=1}^{L} n_k (\bar{z}_k \theta_k - \psi(\theta_k))\right] a^{-(L+g-p)/2} \qquad (5.144) \end{aligned}$$

The result follows now from (5.140). □

To find the posterior distribution of $\boldsymbol{\theta}$ given the data $\boldsymbol{z}$, the direct approach involves high-dimensional numerical integration, and is not computationally feasible. Instead, we discuss here the Gibbs sampling procedure introduced in the earlier sections. Its implementation requires generating samples from the following full conditional distributions. The full conditionals are

(i) $\boldsymbol{b}|\boldsymbol{\theta}, \sigma_u^2 \sim N((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\theta}, \sigma_u^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$;

(ii) $\sigma_u^{-2}|\boldsymbol{\theta}, \boldsymbol{b} \sim gamma((a + \sum_{k=1}^{L}(\theta_k - \boldsymbol{x}_k^T\boldsymbol{b})^2)/2, (L+g)/2)$;

(iii) conditional on $\boldsymbol{b}$ and σ_u^2, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_L$ are mutually independent and

$$\pi(\boldsymbol{\theta}_k|\boldsymbol{b}, \sigma_u^2) \propto \exp[n_k\phi_k^{-1}(\bar{z}_k\theta_k - \psi(\theta_k)) - (\theta_k - \boldsymbol{x}_k^T\boldsymbol{b})^2/2].$$

It is easy to generate samples from the normal and gamma distributions given in (i) and (ii). On the other hand, as evidenced in (iii), the posterior distributions of the θ_k given $\boldsymbol{b}$, and σ_u^2 are known only up to a multiplicative constant. Accordingly, one has to use a general accept-reject algorithm to generate samples from this pdf. Fortunately, the task becomes much simpler due to the following lemma establishing the log-concavity of these posterior pdf's, because one can then use the adaptive rejection sampling scheme of Gilks and Wild (1992) .

Lemma 5.6.1 $\log \pi(\theta_k|\boldsymbol{b}, \sigma_u^2)$ *is a concave function of* θ_k.

Proof.

$$\partial^2 \log \pi(\theta_k|\boldsymbol{b}, \sigma_u^2)/(\partial\theta_k^2) = -n_k\phi_k^{-1}\psi''(\theta_k) - \sigma_u^{-2} < 0,$$

where we use the fact that $\phi_k\psi''(\theta_k) = V(y_k|\boldsymbol{\theta}, \boldsymbol{b}, \sigma_u^2) > 0$. □

Based on the posterior distributions given in (i)–(iii), one can find $E[\theta_k|\boldsymbol{z}]$, $V[\theta_k|\boldsymbol{z}]$ and $Cov[\theta_k, \theta_k'|\boldsymbol{z}]$ using Monte–Carlo integration techniques and iterated conditional expectation and variance formulas.

More general hierarchical models than those given in this section are now available in the literature. Among others, we may refer to Ghosh *et al.* (1996).

A prime issue in any model-based procedure is the choice of appropriate covariates. Inclusion of too many covariates with a high degree of correlation among them will induce multicollinearity, while inclusion of only a few covariates will often lead to a very poor fit. To strike a compromise, a Bayesian approach is to start

with a few plausible models (based on the inclusion or exclusion of certain covariates), say, W with the wth model receiving the prior probability p_w, $w = 1, \ldots, W$. For instance, a default prior will assign a prior probability W^{-1} to each such model. One can compute the posterior probabilities of all the plausible models given the data, and select only those that stand out with high posterior probabilities, or more conservatively eliminate those with low posterior probabilities.

Once an appropriate subset of the W models is selected, an estimation procedure need not be confined to one particular model, nor does one need to provide separate sets of estimates based on each model under consideration. It is possible to provide estimates and associated standard errors based on weighted averages of estimates from all these models, where the weights are determined by the data (e.g. Raftery (1993), Malec and Sedransk (1994), and Draper (1995)). The details for a generalized linear model are worked out in Ghosh *et al.* (1996).

References

Alam, K. (1979). Estimation of multinomial probabilities. *Annals of Statistics*, **7**, 282–283.

Albert, J. (1988). Computational methods using hierarchical Bayes generalized linear model. *Journal of the American Statistical Association*, **83**, 1037–1044.

Andrews, D. F., and Herzberg, A. M. (1985). *A Collection of Problems From Many Fields for the Student and Research Worker.* New York: Springer.

Banks, D. L. (1988). Histospline smoothing the Bayesian bootstrap. *Biometrika*, **75**, 673–684.

Basu, D. (1969). Role of sufficiency and likelihood principles in sample survey theory. *Sankyā B*, **31**, 441–454.

Basu, D. (1971). An essay on the logical foundations of survey, part one. *In:* Godambe, V. P., and Sprott, D. A. (eds), *Foundations of Statistical Inference.* Toronto: Holt, Rinehart and Winston.

Basu, D., and Ghosh, J. K. (1967). Sufficient statistics in sampling from a finite universe. *Bulletin of the International Statistical Institute, BK.2*, **42**, 850–859.

Battese, G., Harter, R., and Fuller, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis.* 2nd edn. New York: Springer-Verlag.

Binder, D. A. (1982). Non-parametric Bayesian models for samples from finite population. *Journal of the Royal Statistical Society, Series B*, **44**, 388–393.

Breslow, N. E., and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brown, L. D. (1981). A complete class theorem for statistical problems with finite sample space. *Annals of Statistics*, **9**, 1289–1300.

Brown, L. D. (1988). Admissibility in discrete and continuous invariant nonparametric estimation problems and in their multinomial analogs. *Annals of Statistics*, **16**, 1567–1593.

Calvin, J., and Sedransk, J. (1991). The patterns of care studies. *Journal*

of the American Statistical Association, **86**, 36–48.

Carter, G., and Rolph, J. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association*, **69**, 880–885.

Chambers, R. L., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, **73**, 597–604.

Chiu, H. Y., and Sedransk, J. (1986). A Bayesian procedure for imputing missing values in sample surveys. *Journal of the American Statistical Association*, **81**, 667–676.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd edn. New York: Wiley.

Cohen, M. P., and Kuo, L. (1985). The admissibility of the empirical distribution function. *Annals of Statistics*, **13**, 262–271.

Datta, G. S., and Ghosh, M. (1991). Bayesian prediction in linear models: applications to small area estimation. *Annals of Statistics*, **19**, 1748–1770.

Dempster, A., Rubin, D. B., and Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, **76**, 341–353.

Dempster, A. P., and Tomberlin, T. J. (1980). The analysis of census undercount from a postenumeration survey. Pages 88–94 of: *Proceedings of the Conference of Census Undercount*. Bureau of the Census.

Diaconis, P., and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, **7**, 269–281.

Doss, H. (1985). Bayesian nonparametric estimation of the median: Part I: Computation of the estimates. *Annals of Statistics*, **13**, 1432–1444.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society, Series B*, **57**, 45–97.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B., and Morris, C. (1973). Stein's estimation rule and its competitors–an empirical Bayes approach. *Journal of the American Statistical Association*, **68**, 117–130.

Ericson, W. A. (1969a). A note on the posterior mean. *Journal of the Royal Statistical Society, Series B*, **31**, 332–334.

Ericson, W. A. (1969b). Subjective Bayesian models in sampling finite populations (with discussion). *Journal of the Royal Statistical Society, Series B*, **31**, 195–233.

Fay, R., and Herriot, R. (1979). Estimates of income for small places: an application of James–Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Fay, R. E. (1987). Application of multivariate regression in small domain estimation. Pages 91–102 of: Platek, R., Rao, J. N. K., Sarndal, C. E., and Singh, M. P. (eds), *Small Area Statistics*. New York: Wiley.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* Vol. I. New York: Wiley.

Ferguson, T. S. (1967). *Mathematical Statistics.* New York: Academic Press.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

Francisco, C. A., and Fuller, W. A. (1991). Quantile estimation with a complex survey design. *Annals of Statistics*, **19**, 454–469.

Gelfand, A., and Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Ghosh, J. K. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays by D. Basu.* New York: Springer-Verlag.

Ghosh, M., and Lahiri, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, **82**, 1153–1162.

Ghosh, M., and Lahiri, P. (1992). A hierarchical Bayes approach to small area estimation with auxiliary information. Pages 107–125 of: Goel, P. K., and Iyengar, N. S. (eds), *Bayesian Analysis in Statistics and Econometrics.* New York: Springer–Verlag.

Ghosh, M., and Meeden, G. (1983). Estimation of the variance in finite population sampling. *Sankyā B*, **45**, 362–375.

Ghosh, M., and Meeden, G. (1986). Empirical Bayes estimation in finite population sampling. *Journal of the American Statistical Association*, **81**, 1058–1062.

Ghosh, M., and Sinha, B. K. (1990). On the consistency between model- and design-based estimators in survey sampling. *Communications in Statistics*, **20**, 689–702.

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. (1996). *Generalized Linear Models for Small Area Estimation.* Technical report 486. University of Florida.

Gilks, W. R., and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics*, **41**, 337–348.

Godambe, V. P. (1966). A new approach to sampling from finite populations, I: Sufficiency and linear estimation. *Journal of the Royal Statistical Society, Series B*, **28**, 310–319.

Godambe, V. P. (1969). Admissibility and Bayes estimation in sampling finite populations V. *Annals of Mathematical Statistics*, **40**, 672–676.

Goldstein, M. (1975). Approximate Bayes solutions to some nonparametric problems. *Annals of Statistics*, **3**, 512–517.

Greenlees, J. S., Reece, W. S., and Zieschang, K. D. (1982). Imputation

of missing values when the response depends on the variable being imputed. *Journal of the American Statistical Association*, **77**, 251–261.

Hartigan, J. A. (1969). Linear Bayesian methods. *Journal of the Royal Statistical Society, Series B*, **31**, 446–454.

Hartley, H. O., and Rao, J. N. K. (1968). A new estimation theory for sample survey. *Biometrika*, **55**, 547–557.

Heath, D., and Sudderth, W. (1976). De Finetti's theorem on exchangeable variables. *The American Statistician*, **30**, 188–189.

Herbach, H. (1959). Properties of Type II analysis of variance tests. *Annals of Mathematical Statistics*, **30**, 939–959.

Hidiroglou, M. A., and Srinath, K. P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, **76**, 690–695.

Hill, B. M. (1968). Posterior distribution of percentiles: Bayes's theorem for sampling from a finite population. *Journal of the American Statistical Association*, **63**, 677–691.

Hsuan, F. C. (1979). A stepwise Bayes procedure. *Annals of Statistics*, **7**, 860–868.

James, W., and Stein, C. (1961). Estimation with quadratic loss. Pages 361–380 of: *Proceedings of the Fourth Berkeley Symposium in Mathematical Statistics and Probability*, V I. University of California Press.

Jessen, R. J. (1978). *Statistical Survey Techniques*. New York: Wiley.

Johnson, B. M. (1971). On admissible estimators for certain fixed sample binomial problems. *Annals of Mathematical Statistics*, **42**, 1579–1587.

Joshi, V. M. (1965). Admissibility and Bayes estimation in sampling finite populations II and III. *Annals of Mathematical Statistics*, **36**, 1723–1742.

Joshi, V. M. (1966). Admissibility and Bayes estimation in sampling finite populations IV. *Annals of Mathematical Statistics*, **37**, 1658–1670.

Kuk, A. Y. C., and Mak, T. K. (1989). Median estimation in the presence of auxiliary information. *Journal of the Royal Statistical Society, Series B*, **51**, 261–269.

Lahiri, P. (1986). *Robust Empirical Bayes Estimation in Finite Population Sampling*. Ph.D. thesis, University of Florida.

Lane, D. A., and Sudderth, W. D. (1978). Diffuse models for sampling and predictive inference. *Annals of Statistics*, **6**, 1318–1336.

Lehmann, E. L., and Scheffe, H. (1950). Completeness, similar regions and unbiased estimation. *Sankyā B*, **10**, 305–340.

Leonard, T. (1976). Some alternative approaches to multiparameter estimation. *Biometrika*, **63**, 69–76.

Leonard, T., and Novick, M. R. (1986). Bayesian full rank marginalization for two-way contingency tables. *Journal of Educational Statistics*,

11, 33–56.

Lindley, D. (1962). Discussion of Professor's Stein's paper 'Confidence sets for the mean of a multivariate normal distribution'. *Journal of the Royal Statistical Society, Series B*, **24**, 265–296.

Lindley, D., and Smith, A. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

Little, R. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, **78**, 596–604.

Lo, A. Y. (1988). A Bayesian bootstrap for a finite population. *Annals of Statistics*, **16**, 1684–1695.

MacGibbon, B., and Tomberlin, T. J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology*, **15**, 237–252.

Malec, D., and Sedransk, J. (1985). Bayesian inference for finite population parameters in multistage cluster sampling. *Journal of the American Statistical Association*, **80**, 897–902.

Malec, D., and Sedransk, J. (1994). *Small area inference for binary variables in the National Health Interview Survey.* Technical report SUNY, Albany.

Malec, D., Sedransk, J., and Tomkins, L. (1993). Bayesian predictive inference for small areas for binary variables in the National Health Interview Survey. Pages 377–389 of: Gatsonis, C., Hodges, J. .S, Kass, R. E., and Singpurwalla, N. D. (eds), *Case Studies in Bayesian Statistics.* New York: Springer-Verlag.

Mazloum, R., and Meeden, G. (1987). Using the stepwise Bayes technique to choose between experiments. *Annals of Statistics*, **15**, 269–277.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd edn. London: Chapman & Hall.

Meeden, G. (1990). Admissible contour credible sets. *Statistics and Decisions*, **8**, 1–10.

Meeden, G. (1992). The admissibility of the linear interpolation estimator of the population total. *Annals of Statistics*, **20**, 510–522.

Meeden, G. (1993). Noninformative nonparametric Bayesian estimation of quantiles. *Statistics & Probability Letters*, **16**, 103–109.

Meeden, G., and Bryan, M. (1996). An approach to the problem of nonresponse in sample survey using the Polya posterior. Pages 423–431 of: *Bayesian Analysis in Statistics and Econometrics Essays in Honor of Arnold Zellner.* New York: Wiley.

Meeden, G., and Ghosh, M. (1981a). Admissibility in finite problems. *Annals of Statistics*, **9**, 846–852.

Meeden, G., and Ghosh, M. (1981b). On the admissibility and uniform admissibility of ratio type estimators. Pages 378–390 of: *Statistics:*

Applications and New Direction. Indian Statistical Institute Golden Jubilee International Conference.

Meeden, G., and Ghosh, M. (1983). Chosing between experiments: applications to finite population sampling. *Annals of Statistics*, **11**, 296–305.

Meeden, G., and Vardeman, S. (1985). Bayes and admissible set estimation. *Journal of the American Statistical Association*, **80**, 465–471.

Meeden, G., Ghosh, M., and Vardeman, S. (1985). Some admissible nonparametric and related finite population sampling estimators. *Annals of Statistics*, **13**, 811–817.

Meeden, G., Ghosh, M., Srinivasan, C., and Vardeman, S. (1989). The admissibility of the Kaplan–Meier and other maximum likelihood estimators in the presence of censoring. *Annals of Statistics*, **17**, 1509–1531.

Morris, C. (1982). Natural exponential families with quadratic variance functions. *Annals of Statistics*, **11**, 65–80.

Morris, C. (1983a). Natural exponential families with quadratic variance functions: statistical theory. *Annals of Statistics*, **11**, 515–529.

Morris, C. (1983b). Parametric empirical Bayes confidence intervals. Pages 25–50 of: Box, G. E. P., Leonard, T., and Wu, C. F. J. (eds), *Scientific Inference, Data Analysis and Robustness.* New York: Academic Press.

Morris, C. (1983c). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, **78**, 47–54.

Muliere, P., and Secchi, P. (1996). Bayesian nonparametric predictive inference and bootstrap techniques. *Annals of the Institute of Statistical Mathematics*, **48**, 663–673.

Prasad, N. G. N., and Rao, J. N. K. (1990). On the estimation of mean square error of small area predictors. *Journal of the American Statistical Association*, **85**, 163–171.

Raftery, A. (1993). *Approximate Bayes factors and accounting for model uncertainty in generalized linear models.* Technical report 255. University of Washington.

Raiffa, H., and Schlaiffer, R. (1961). *Applied Statistical Decision Theory.* Boston: Harvard University Press.

Raj, D. (1968). *Sampling Theory.* New York: McGraw-Hill.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications.* New York: Wiley.

Rao, J. N. K., Kovar, J. G., and Mantel, H. J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, **77**, 365–375.

Robbins, H. (1956). An empirical Bayes approach to statistics. Pages 157–163 of: *Proceedings of the Third Berkeley Symposium in Mathe-*

matical Statistics and Probability, V I. University of California Press.

Robbins, H. (1983). Some thoughts on empirical Bayes estimation. *Annals of Statistics*, **11**, 713–723.

Royall, R. (1971). Linear regression models in finite population sampling theory. Pages 259–274 of: Godambe, V. P., and Sprott, D. A. (eds), *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Royall, R. (1976). The least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, **71**, 657–664.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, **57**, 377–387.

Royall, R. M., and Cumberland, W. D. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, **76**, 66–88.

Royall, R. M., and Cumberland, W. D. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*, **80**, 355–359.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, **72**, 538–543.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics*, **9**, 130–134.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scott, A. J. (1975). On admissibility and uniform admissibility in finite population sampling. *Annals of Statistics*, **3**, 489–491.

Scott, A. J. (1977). On the problem of randomization in survey sampling. *Sankyā B*, **39**, 1–9.

Scott, A. J., and Smith, T. M. F. (1969). Estimation in multistage surveys. *Journal of the American Statistical Association*, **64**, 830–840.

Searle, S. (1971). *Linear Models*. New York: Wiley.

Sengupta, S. (1980). On the admissibility of the symmetrized Des Raj estimator for ppswor samples of size two. *Calcutta Statistical Association Bulletin*, **29**, 35–44.

Smith, P. J., and Sedransk, J. (1982). Bayesian optimization of the estimation of the age composition of a fish population. *Journal of the American Statistical Association*, **77**, 707–713.

Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Annals of the Institute of Statistical Mathematics*, **16**, 155–160.

Stroud, T. W. F. (1987). Bayes and empirical Bayes approaches to small area estimation. Pages 124–137 of: Platek, R., Rao, J. N. K., Sarndal, C. E., and Singh, M. P. (eds), *Small Area Statistics.* New York: Wiley.

Stroud, T. W. F. (1991). Hierarchical Bayes predictive means and variances with application to sample survey inference. *Communications in Statistics*, **20**, 13–36.

Stroud, T. W. F. (1994). Bayesian inference from categorical survey data. *Canadian Journal of Statistics*, **22**, 13–36.

Sugden, R. A., and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, **71**, 495–506.

Vardeman, S., and Meeden, G. (1983). Admissible estimators in finite population sampling employing various types of prior information. *Journal of Statistical Planning and Inference*, **7**, 329–341.

Vardeman, S., and Meeden, G. (1984). Admissible estimators for the total of a stratified population that employ prior information. *Annals of Statistics*, **12**, 675–684.

Wald, A., and Wolfowitz, J. (1951). Characterization of the minimal complete class of decision functions when the number of distributions and decisions is finite. Pages 149–157 of: *Proceedings of the Second Berkeley Symposium in Mathematical Statistics and Probability*, V I. University of California Press.

Wilks, S. S. (1962). *Mathematical Statistics.* New York: Wiley.

Woodruff, R. S. (1952). Confidence intervals for the median and other position measures. *Journal of the American Statistical Association*, **47**, 635–646.

Yang, S.-S. (1985). A smooth nonparametric estimator of a quantile function. *Journal of the American Statistical Association*, **80**, 1004–1011.

Zacks, S. (1969). Bayesian sequential designs for sampling finite populations. *Journal of the American Statistical Association*, **64**, 1342–1349.

Zeger, S. L., and Karim, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, **86**, 79–86.

Author index

Subject index