USE OF THESES

This copy is supplied for purposes of private study and research only. Passages from the thesis may not be copied or closely paraphrased without the written consent of the author.
CORRECTIONS AND EXPLANATIONS

p34 line 9. For "population" read "population unit".

p36, 2nd last line. For "estmator" read "estimator".

p39 line 3. $x$ is the population vector of a "benchmark" (or "supplementary" or "auxiliary") variable, $x$. For a ratio estimator, such a variable must be
(a) known for every sample unit and for the population total and
(b) sufficiently closely proportional to the survey variable $y$ for $T(x)/\hat{T}_{HT}(x)$ to be a useful correction factor to apply to $\hat{T}(y)$.

p40 line 10. $X$ is a matrix of "benchmark" variables (see entry above). For a regression estimator, a benchmark variable need not be proportional to the survey variable, but it must have some explanatory power. For some purposes it is a requirement that every element of $X$ be known, for instance if we wish to select a sample that is balanced of order $L > 1$ (see p45). Often, however, only its column sums are known. These column sums themselves are sometimes called "benchmarks."

p41 line 7. For $n^{-1}$ read $Nn^{-1/2}$. The argument here needs clarification From p40, $\hat{T}_{GREG}(y) = \hat{T}_{HT}(y) + \{T(X) - \hat{T}_{HT}(X)\}\hat{\beta}_{GREG}$ where $\hat{\beta}_{GREG} = (W'X)^{-1}W'y$. It is of course best if the $W_s$ are such that $\hat{\beta}_{GREG}$ estimates $\beta$ reasonably closely, but since $\hat{T}_{GREG}(y) = [\hat{T}_{HT}(y) + \{T(X) - \hat{T}_{HT}(X)\}\beta] + [(T(X) - \hat{T}_{HT}(X))\beta - \hat{\beta}_{GREG}]$, and since any definition of $\hat{\beta}_{GREG}$ that converges to $\beta$ as $n^{-1/2}$ ensures that the second expression in square brackets will be of order $n^{-1}$, the precise choice of the $W_s$ is usually of little consequence.

p45 last line of footnote. For "article" read "thesis".

p48 line 7. For a definition of a stratum-by stratum estimator, see the displayed equation on p52 (line 8).

p52 2nd last line. Delete closing bracket.

p53 line 9. For "its model-expected value" read "the model-expectation of that bias".

p61 line 6. The expression $w_{ij}$ denotes the "case weight" for the $j$th unit within the $h$th stratum. The most general sample estimator of $T(y_h)$ can therefore be written
\[ \hat{T}(y_h) = \sum_{j \in s_h} w_{ij} y_{ij}. \] Further, the use of the ratio estimator $\left( \sum_{j \in s_h} y_{ij} / \sum_{j \in s_h} x_{ij} \right) T(x_h)$ with a
sample balanced on $x$ implies $w_{hj} = \left( \frac{T(x_h)}{\sum_{j \in s_h} x_{hk}} \right) = N_h / n_h$ for all $j \in s_h$. The selection of a simple random sample instead, and the calibration of the estimator $\hat{T}(y_h)$ to ensure model-unbiasedness for $T(y_h)$, will in general lead to $w_{hj}$ values that differ from $N_h / n_h$. For any variable $y$ unrelated to $x$ and obeying instead the model

$$\xi_h \quad y_j = \mu + \epsilon_j \text{ where } E_{\xi_h} \epsilon_{hj} = 0, \quad E_{\xi_h} \epsilon_{hj}^2 = \sigma^2 \quad \text{and} \quad E_{\xi_h} \epsilon_h \epsilon_{hk} = 0, \text{ all } k \neq j,$$

the model-variance of $\bar{n}_h$ is $\sigma^2 \sum_{j \in s_h} w_{hj} (w_{hj} - 1)$. The effective sample size $\bar{n}_h$ is then defined implicitly by the relationship $V(\hat{T}(y_h)) = \sigma^2 N_h^2 / \bar{n}_h$. If $w_{hj} = N_h / n_h$ for all $j \in s_h$, $\bar{n}_h = n_h N_h / (N_h - n_h)$, the factor $N_h / (N_h - n_h)$ being the reciprocal of the finite population correction factor. For general $w_{hj}$, however, $\bar{n}_h = N_h^2 / \sum_{j \in s_h} w_{hj} (w_{hj} - 1)$.

p68 line 9. For $\{y_j - x'_{\beta} \} \pi_i^{-1}$ read $\{y_k - x'_{\beta} \} \pi_k^{-1}$.

p71 line 13. For "Section 6" read "Section 8 of this Chapter".

p75 last line. The four nuances of interpretation referred to are those on p76, namely (i) Cosmetic Estimation, (ii) Cosmetic Prediction, (iii) & (iv) Calibration Estimation and (v) Calibration Prediction. There is arguably a fifth nuance if the short and long forms of the Calibration Estimator are considered separately, the long form more nearly indicating its relationship to the Calibration Predictor.

p78 line 7. For "is less" read "would usually be less".

p78 lines 9-10. When each of the $w_j$ is equal to unity, each sample unit represents itself only. This usually occurs when the population has been completely enumerated. (If the sample size is fixed, no alternative is possible.) For complete enumerations, the model-variance $E_{\xi} (\hat{T}(y) - T(y))^2$ should be, and usually is, zero.

p78 4th last line. Delete the unmatched {.

p82 line 16. For "also" read "necessarily".

p86 lines 12 and 13. For 45 read 46.

p86 line 13. For "seven" read "six".

p88 last line of Table A.2. For $>1.4$ read $\geq 1.2$. 
1. ERROR ON P71.

The last paragraph is in error because \( (Z^{-1}X) \) is not diagonal and so cannot have an inverse. A correct proof would proceed as follows.

To this end, we replace \( A^{-2} \) in the expression \( \hat{\beta}_{BLUE} = [X'\mathbf{A}^{-2}X]^{-1}X'y \) by \( Z^{-1}_s(\Pi^{-1}_s-I_n) \) where \( Z_s \) is an \( n \times n \) diagonal matrix, \( Z_s1_n \) being any linear function of the columns of \( X_s \). Then \( \hat{\beta}_{cos} = [X_s'Z_s^{-1}(\Pi^{-1}_s-I_n)X_s]^{-1}X_s'Z_s^{-1}(\Pi^{-1}_s-I_n)y_s \), \( Q_s = (\Pi^{-1}_s-I_n)Z_s^{-1}X_s \), and \( \alpha_s = (X_s'X_s)^{-1}X_s'Z_s^{-1}X_s. \) It follows that \( Q_s\alpha_s = (\Pi^{-1}_s-I_n)1_n \) as required.
Reconciling some apparently incompatible approaches to statistical inference

A thesis submitted for the degree of
Doctor of Philosophy
of the Australian National University
December 1996

Kenneth Ronald Walter Brewer
STATEMENT OF ORIGINALITY

I hereby certify that this thesis does not contain any material previously published or written by any other person except where due reference is made in the text.

K.R.W. Brewer
TOPIC B: FREQUENTIST AND BAYESIAN TESTS OF A PRECISE NULL HYPOTHESIS AGAINST A COMPOSITE ALTERNATIVE

Prof. R.L. Chambers drew my attention to the challenge of Stone's (1976) Treasure Hunt in Flatland. Prof. D.V. Lindley provided useful and encouraging comments on early drafts of Part I. Dr Alan Dorfman, in a seminal discussion, helped to pull the ideas in it together.

Prof. Murray Aitkin's seminars on the Posterior Bayes Factor triggered my initial resolve to find an acceptable Reference Bayesian test of a precise null hypothesis against a composite alternative. This provided the focus for Parts II and III. Dr Don Poskitt, Dr Robert Smythe and Mr John Morgan provided early encouragement. Prof. D.V. Lindley and Prof. J.O. Berger pointed out some serious pitfalls along the way.
To the memory of my father and mother
Abstract

This thesis consists of an Introduction and two Topics. The Introduction deals with the underlying theme of the thesis and why these two particular topics were chosen to illustrate it. It is argued that the contradictions we meet in statistics (as in the physical sciences) may only appear to be such as a result of our imperfect understanding, and that in consequence it is often healthy to hold them in a creative tension.

Topic A deals with the design-based and model-based approaches to sampling inference in the context of business surveys. Each has its strong advocates and those who hold to either view typically regard the other as misguided and worthless. It is shown that both have their advantages and disadvantages; also that it is possible to devise schemes which enable both to be used at the same time, reaping nearly all the advantages and avoiding nearly all the disadvantages of both, and obtaining new insights and options into the bargain. Part I is almost purely expository, but Parts II and III indicate how the synthesis can be achieved, in the contexts of stratified and unequal probability designs respectively. Part III includes an empirical study. Part IV provides an assessment of what can be achieved by using the two approaches simultaneously.

Topic B deals principally with the disparate approaches used by Frequentist and Bayesian statisticians to the testing of a precise null hypothesis against a composite alternative. It is shown that starting from identical evidence they can easily come to opposite conclusions as to the choice that should be made between the two, particularly when the problem is ill-specified, but also even when it is not. The author was trained as a Frequentist, and although now leaning to the view that Bayesian inference, when correctly applied, can usually supply a deeper insight, he is acutely aware of certain areas where current Bayesian practice is wholly inadequate. The current Bayesian procedure for testing a precise null hypothesis against a composite alternative is the most remarkable instance of this failure, and it turns out that when a more robust Bayesian test is specified, the relevant test statistic is closely
related to that already used by Frequentists. It is argued that, properly understood, the two approaches are both valid, and each can provide useful insights for the other.

The hypothesis testing problem is here approached via Stone's (1976) notorious Treasure Hunt in Flatland, which caught his Bayesian contemporaries sadly off-balance. It transpires that there is actually a better solution to Stone's problem than the one used by his Classical Statistician, and that it can be found straightforwardly using "empirical Bayes." A fortunate ambiguity in Stone's own diagram enables both the strengths and the weaknesses of Bayesian inference to be exhibited. If (as in Part I) the Treasure Layers followed a clockwise path through a certain loop, the relevant hypothesis test is between two precise hypotheses, a task that current Bayesian practice can handle quite readily. If, however, the path they followed included an anticlockwise loop (as in Part II), the test is between a precise null and a potentially diffuse alternative. Current Bayesian practice in this situation requires a subjective specification of the alternative prior, and it is shown that this can lead to very different answers, some of which are highly counterintuitive. The properties of the newly specified Bayesian test statistic are considered in Part III, and this is where the relationship with Frequentist inference is discovered. Once again, the conclusion is that the apparent contradictions are best held in creative tension.
RELATED PUBLICATIONS

Papers based on the results of this thesis already published or submitted for publication are


## CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement of Originality</td>
<td>i</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Abstract</td>
<td>v</td>
</tr>
<tr>
<td>Related Publications</td>
<td>vii</td>
</tr>
<tr>
<td>Contents</td>
<td>viii</td>
</tr>
<tr>
<td>INTRODUCTION: Contradictions and Creative Tension</td>
<td>1</td>
</tr>
<tr>
<td>1. Contradictions in the Physical Sciences</td>
<td>1</td>
</tr>
<tr>
<td>2. Contradictions in Statistical Inference</td>
<td>3</td>
</tr>
<tr>
<td>Topic A: Design-based and Model-based Inference in Business Survey Sampling</td>
<td>3</td>
</tr>
<tr>
<td>Topic B: Frequentist and Bayesian Tests of a Precise Null Hypothesis against a Composite Alternative</td>
<td>10</td>
</tr>
<tr>
<td>3. The Need for Creative Tension</td>
<td>20</td>
</tr>
<tr>
<td>References</td>
<td>25</td>
</tr>
<tr>
<td>TOPIC A: Design-based and Model-based Inference in Business Survey Sampling</td>
<td>29</td>
</tr>
<tr>
<td>Part I: Understanding the Differences</td>
<td>30</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>31</td>
</tr>
<tr>
<td>2. Ordinary Design-based Inference</td>
<td>31</td>
</tr>
<tr>
<td>3. Model Assisted Survey Sampling</td>
<td>39</td>
</tr>
<tr>
<td>4. Model-based Inference</td>
<td>41</td>
</tr>
<tr>
<td>Part II: Stratified Random and Stratified Balanced Sampling</td>
<td>51</td>
</tr>
<tr>
<td>1. Where Balanced Sampling Has the Edge</td>
<td>52</td>
</tr>
<tr>
<td>2. Where Randomized Sampling Has the Edge</td>
<td>57</td>
</tr>
<tr>
<td>3. Calibration: Getting the Best of Both Worlds</td>
<td>59</td>
</tr>
<tr>
<td>4. Summary</td>
<td>62</td>
</tr>
<tr>
<td>TOPIC A: Continued</td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---</td>
</tr>
<tr>
<td>Part III: Unequal Probability Sampling</td>
<td>64</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>65</td>
</tr>
<tr>
<td>2. Design-based and Anticipated Variance</td>
<td>68</td>
</tr>
<tr>
<td>3. The Cosmetic Approach to Combined Estimation</td>
<td>70</td>
</tr>
<tr>
<td>4. The Calibration Approach to Combined Estimation</td>
<td>74</td>
</tr>
<tr>
<td>5. Comparing the Cosmetic and Calibration Approaches</td>
<td>75</td>
</tr>
<tr>
<td>6. The Model-variance of $\hat{\tau}_{coscal}(y)$ and its Estimation</td>
<td>77</td>
</tr>
<tr>
<td>7. The Problem of Unacceptably Small Sample Weights</td>
<td>79</td>
</tr>
<tr>
<td>8. An Empirical Study Using Farm Data</td>
<td>80</td>
</tr>
<tr>
<td>Part IV: Some Final Comments</td>
<td>93</td>
</tr>
<tr>
<td>References</td>
<td>100</td>
</tr>
</tbody>
</table>

| TOPIC B: Frequentist and Bayesian Tests of a Precise Null Hypothesis against a Composite Alternative | 106 |
| Part I: An Empirical Treasure Hunt in Flatland | 107 |
| 1. Introduction | 108 |
| 2. The Classical Solution Revisited | 111 |
| 3. Reproducing the Classical Solution Using Bayesian Arguments | 115 |
| 4. An Empirical Bayesian Approach | 117 |
| 5. The Clockwise Scenario | 119 |
| 6. Concluding Remarks | 128 |
| Part II: The Anticlockwise Scenario | 130 |
| 1. The Challenge of the Anticlockwise Scenario | 131 |
| 2. The Reference Prior Problem | 133 |
| 2.1 The Normal Distribution | 134 |
| 2.2 The Binomial Distribution | 142 |
| 3. The Solution for the Anticlockwise Scenario | 146 |
| 4. Concluding Remarks | 149 |
TOPIC B: Continued

Part III: Properties of the Reference Posterior Odds 152
1. Introduction 153
2. The Bayes Factor and its Expectation 155
3. The Reference Posterior Odds 157
   Example 3.1 Normal Mean, Variance Known 160
   Example 3.2 Normal Variance, Mean Unknown 161
   Example 3.3 Binomial Parameter 163
4. The Posterior Bayes Factor 165
5. The Classical or Frequentist Test Statistic 169
6. The Choice of Information Criterion 171
7. The Bayesian-Frequentist Interface 172

References 174
Introduction: Contradictions and Creative Tension

Alice laughed. "There's no use trying," she said. "One can't believe impossible things."

"I dare say you haven't had much practice," said the [White] Queen. "When I was your age, I always did it for half an hour a day. Why, sometimes I've believed as many as six impossible things before breakfast."

Lewis Carroll, *Through the Looking Glass*, Chapter V.

1. CONTRADICTIONS IN THE PHYSICAL SCIENCES

Few things in life can be more impossible than contradictions, yet some kinds of people (and scientists in particular) do live with them from time to time, frequently because they have little choice. A century ago, the sciences of palaeontology and astronomy held contradictory views on the age of the earth. Palaeontologists believed that their observations pointed to changes in extinct life forms that needed hundreds of millions of years to account for them. Geologists supported them in requiring this kind of time scale. Astronomers believed equally firmly that, since the sun had no energy source that could have kept it hot enough to sustain life for more than 20 million years, the palaeontologists and geologists could not possibly be right and that some day they would find some mechanism that would speed things up for them.

In fact, the opposite happened. The astronomers had been explaining the sun's radiation "on the hypothesis of a gaseous body maintaining its temperature by its internal heat and obeying the laws of gases as known to
terrestrial experiment." It was to take several decades before they had enough knowledge to be able to see the sun as consisting of plasma rather than gas and as drawing its energy from the fusion of hydrogen into helium. But they made numerous important discoveries in the meantime. People with wrong ideas can still produce useful results.

One may wonder, perhaps, what would have happened had astronomers taken the trouble to examine the palaeontological and geological evidence carefully. Might they have become convinced, and looked harder for some other energy source in the sun? Might the discoveries of atomic and nuclear physics been speeded up in consequence? But then, what if palaeontologists and geologists had looked carefully at the astronomical evidence? Might they have wasted endless hours trying to find some non-existent mechanism that would speed up their own time scale? Perhaps what actually happened was for the best after all. Both groups of scientists were convinced that they and they only were right and each continued to ignore the other, confident of eventual vindication.

Not all contradictions in science, however, have been resolved simply by the discovery that one or the other view was demonstrably incorrect, leaving the other in undisputed possession of the field. Nor are they likely to be. It is notoriously impossible to explain all optical phenomena either exclusively in terms of waves or exclusively in terms of particles. We may be able to use mathematically based theory to predict what will be observed in any particular experiment, but we are not yet(?) able to use our imaginations to picture even such helpful images as "particles guided by waves" without running up
against contradictions. As far as I am able to understand, working physicists manage to cope by holding such incompatible concepts in a kind of creative tension.

2. CONTRADICTIONS IN STATISTICAL INFERENCE

By comparison with the contradictions that physical scientists have to accommodate, those faced by statisticians are quite trivial. We seldom disagree fundamentally about what is the case. We typically have only incomplete data about a population, but we agree as to what those data are. We also agree that we can make no statements that are both certain and precise about the population as a whole—unless and until we can observe the rest of it, which is often impossible. All that we disagree about is the usefulness or otherwise of different kinds of inference from the known part to the unknown part. Yet, curiously, we who pride ourselves as experts on uncertainty are often very certain in our own minds that one or another form of statistical inference is "valid" to the exclusion of any other. Can we not, like the optical physicists, hold these ideas in creative tension?

TOPIC A: Design-based and Model-based Inference in Business Survey Sampling

The area of statistics in which I first came up against apparently contradictory theories of inference was survey sampling, also known as finite population sampling. I started to work in this field under E.K. Foreman in 1954. At that time, Neyman's (1934) randomization paradigm was dominant. Hansen and
Hurwitz had used it to remarkably good effect, both in theory and in practice. Of their 1943 paper, Smith (1994) had to say,

"[W]e are presented not just with a theoretical extension of the work of Neyman but a completely new product. For the first time we have a complete theory for stratified multi-stage surveys covering random design, practical implementation analysis and interpretation within the philosophical framework of the randomization distribution. It was the complete package that was so novel, and the whole was so much greater than the sum of the parts. Where the U.S. Census Bureau led others followed, and within a few years almost all major surveys carried out by governments had adopted these ideas."

Foreman had brought this "complete package" back to Australia with him in 1953, and I was privileged to be one of his implementation team almost from the start. Together with all my colleagues, I absorbed without question the randomization orthodoxy on which it was founded, and can remember vividly the scorn with which we then regarded all surveys that were not based on probability samples, for there were still a few around in those days.

I can also remember equally vividly the first occasion on which the randomization paradigm failed to provide me with an insight that I was seeking. The context was one of the Australian Bureau of Statistics' Business Surveys. A single business in a completely enumerated stratum had failed to respond. As was customary, I calculated a separate ratio estimate for that stratum as though a random sample of $N-1$ units had been selected from the $N$
in total. The estimate made good sense. I also calculated, as was equally customary, an estimate of its standard error, based on the same assumption. That estimate did not make such good sense. It so happened that one of the businesses in that stratum was enormously large. The assumption that it was just as likely not to be included in the "random" sample as any of the others led to a completely unrealistic estimate of variance.

Were I faced with the same problem today, I might divide that particular stratum into two—the other businesses, from which non-response could be tolerated, and that particular one from which it could certainly not be tolerated. I would then obtain a more reasonable estimate of the "variance". At that time, however, my experience and understanding were somewhat limited. If I remember rightly, I simply treated that particular ratio estimate as being without sample error, arguing to myself, with some justification, that it was not "really" a sampling error at all, but a non-response error, and could therefore be omitted. It nevertheless troubled me that the ratio estimation procedure had supplied a sensible estimate of total but a completely nonsensical estimate of its variance.

Various other questions that the randomization paradigm seemed unable to answer cropped up from time to time, and I cannot now remember the exact sequence. One related to the relationship between stratified random sampling and unequal probability sampling. In the former, the optimal allocation of sample to strata was well known from Neyman (1934) [though discovered earlier by Tschuprow (1923)] to be governed by the rule that the sampling fraction, $n/N$, should be proportional to the relevant population standard
deviation, or at least to $\sqrt{N/(N-1)}$ times that population standard deviation, a quantity commonly denoted by $S$.

Where there were only two or three size strata, there were typically large differences from stratum to stratum in the values of $S$, so the allocation was some distance away from being optimal. However, as the strata were made more numerous, and hence smaller, $S$ became a less and less stable quantity. In the limit, each stratum would have consisted of a single unit and the values of $S$ would have been indeterminate. I wanted to understand what the optimum inclusion probabilities would be in that situation, and the randomization paradigm in isolation could give me no answer. It did not even permit the question to be framed meaningfully. What could "optimum" mean in that context, with every stratum requiring representation in sample, and with the minimum and maximum sample sizes in each stratum both being unity?

Another problem arose when I was required to design a sample of businesses that manufactured men's outerwear. This was heavily dominated by a few large clothing companies, but there were also many small tailoring businesses that made clothes to measure. Everywhere within the feasible range of sample size, $n$, the Neyman-optimal sample allocations were clearly less efficient than the "cut-off sample" or "partial collection" made up of the largest $n$ units. What was the variance of the ratio estimator of total in such circumstances? Within the randomization paradigm the only meaningful definition of variance involved comparing all possible samples, and here there was only a single possible sample. Once again, there was no question that could be posed in a meaningful fashion.
The final straw came in 1962. I had been working with methods of selecting samples with unequal probabilities without replacement. The two variance formulae for the Horvitz-Thompson estimator were both dependent on the second order inclusion probabilities, and so were the formulae for the estimators of variance (Horvitz and Thompson 1952, Sen 1953, Yates and Grundy 1953). Moreover, the influence on the variance from those second order inclusion probabilities was uncomfortably large. I felt instinctively that the variance depended almost exclusively on the first order inclusion probabilities. I began to wonder under what circumstances the second order ones might be seen as irrelevant.

Another obvious and closely related question to ask was, "When the strategy of selecting with unequal probability with replacement using the Hansen-Hurwitz (1943) estimator is replaced by the strategy of selecting with unequal probability without replacement using the Horvitz-Thompson (1952) estimator, by how much is the variance reduced?"

As I pondered these last two questions, it became clear that all five could be answered if I confined my attention to the particular situation where the survey variable values were generated by a certain stochastic model. It was not a new model. I was already familiar with it from Cochran (1953), although he seemed there to be using it with a slightly apologetic air. Later, I traced its use back past Godambe (1955) and Jessen (1942) to H.F. Smith (1938), each of whom had used it to solve a single problem.
The immediate result was Brewer (1963), a paper that I look back to with mixed feelings. To the extent that I saw the use of this model as a key to obtain answers to many questions that could not even be formulated within the randomization paradigm, it contained a real insight; but it was poorly written and its impact outside the Australian Bureau of Statistics was close to zero. I attempted to sell my idea to M.H. Hansen during my five-month visit to the U.S. Census Bureau in 1966-67, but was unable to spark any interest. For him, models were a snare and a delusion that we knew better than to have any further truck with.

I had more success with Foreman. The limited role for models found in Foreman and Brewer (1971) and later in Brewer, Foreman, Mellor and Trewin (1977) was the same as what is now known as model assisted survey sampling (Särndal, Swensson and Wretman 1992). On the one hand these authors were to set out (page v) "[t]o write a basic sampling text that, unlike its predecessors, is guided by statistical modelling in the derivation of estimators." On the other hand, they were also to state (page vi), "We use the randomization theory, or design-based, point of view. This is the traditional mode of inference in surveys, ever since the sampling theory breakthroughs in the 1930s and 1940s. The reasoning is familiar to survey statisticians in government and elsewhere."

But all that was then a long way ahead. More immediately, Royall's (1970, 1971) vigorous attack on the randomization paradigm as an unsuitable vehicle for statistical inference was to set the survey sampling world by its ears. Shortly afterwards, Royall and Herson (1973a,b) were to supplement this with
the idea of balanced sampling and provide a fully viable alternative paradigm based on modelling, with the H.F. Smith (1938) model occupying a central position. I had earlier shown (Brewer 1963) how model-based inference could be used in the context of a "partial collection", but had cautioned that in more usual circumstances, "it might well be unwise to abandon a sampling plan for which a variance can be calculated, regardless of any assumptions, for one giving only a conditional variance valid on the assumption that [the model equations] describe the generation of the population." After 1973, I was less worried about that risk and started to think of the model-oriented paradigm as a serious alternative to randomization.

The next stimulus to my thinking about these issues came when I attended the ISI Congress in New Delhi in 1977 and heard Godambe say that what we now needed were estimators that made sense in both approaches at the same time. From then on, I was determined to pursue that enterprise. It was to be far more difficult than I had supposed. Brewer (1979) was an important step in that direction and led me to propound ideas for design-consistency and asymptotic design-unbiasedness that attracted sufficient interest for various people to suggest slightly modified versions of it (see Topic A, Part III). But progress was slow. Brewer, Hanif and Tam (1988) were still asking the question as to how far such integration was possible.

In 1992, I started to write another paper on the topic. It was to be presented at the International Conference on Establishment Surveys in Buffalo and, if all went well, published in a monograph. Several people, notably Prof R.L. Chambers (then my PhD supervisor), Dr D.A. Binder (my Section Editor)
and Dr P.S. Kott (who took a particular interest in my paper on Binder's behalf) supplied me with many stimulating comments to keep me on the rails. Chambers emphasized from the beginning the importance of case-weights. (A case-weight is a sample weight unique to a given sample unit, to be applied to all survey variables alike.) Binder was very concerned about the direction and construction of the paper, while Kott hammered away at the details. By the time my paper was delivered in June 1993, I believed I had succeeded in meeting Godambe's 1977 challenge. Brewer (1995) is the written version of that paper. Shortly afterwards, however, Chambers tried it out and found it wanting. It looked good in theory but was disappointing in practice. Fortunately, it only required fine tuning to get it to work properly (Brewer 1994 and Topic A, Part III). Part I is an introduction to the whole of Topic A that draws largely on Brewer (1995). Part II represents the plugging of a gap: material that would have formed part of Brewer (1995) but for limitations of time and page space. Part IV supplies a summary and an assessment of the relevance of the other three Parts.

**TOPIC B: Frequentist and Bayesian Tests of a Precise Null Hypothesis against a Composite Alternative**

Though brought up as a Frequentist, I have long regarded the logic of the Bayesian approach as attractive. I have also been uneasy with the counterintuitive relevance of optional stopping rules in Frequentist inference, and even more with the language that Frequentists use to describe their constructs. One of their apologists, Alexander (1994), has ventured to explain confidence intervals as follows:
"'Keep in mind that the particular sample we selected was one of many possible samples. ... Different samples would give different results.

"'If many samples were selected, then for approximately 95% of the samples the confidence interval which would be calculated would contain the result which would be obtained from surveying the entire population.'

"It sounds like someone's worst nightmare from Freshman Statistics, [he comments,] but it does two things:

(a) It gives the reader a concrete image to reinforce the notion that there is uncertainty because of sampling error.

(b) It precisely and completely states the fact upon which we expect the readers to base their statistical inferences about sampling error ...

"I have to admit that what people would really like to know is the probability that the population value is in the confidence interval, which isn't exactly what we non-Bayesians can tell them."

In this short passage, Alexander exposes the two major difficulties that Frequentists have with their public relations; namely that "people" find their concepts difficult to follow, and that if they do take the trouble to understand
them, they often decide that they are not what they really want. Over four decades of a statistical career, I have found three typical attitudes towards confidence intervals among non-specialists. Using the expression "population value" to denote that which would be found if every one of the population units were measured instead of only a sample, I can describe these three attitudes as follows:

1. Naive Fiducialism: "If it's 19 to one on that a 95 per cent confidence interval will contain the population value, it must be 19 to one on that this particular 95 per cent confidence interval contains the population value."

2. Science-blindness: "I don't really understand what they mean by all this confidence interval business, but they are the experts, so it must be all right to use them."

3. Resigned Pragmatism: "I understand what they are saying well enough to know that confidence intervals aren't telling me what I really want to know, but everyone else uses them and nothing appalling ever seems to happen, so I shall almost certainly get by if I do the same."

I would have preferred instead to produce Bayesian credibility intervals (almost invariably using uniform priors) and explain them as follows:

"On the assumption that, before the sample measurements were taken, we were completely ignorant about the population value, the odds are
now 19 to one on that it lies within the credibility interval we have specified."

But I never even tried to do so. What kept me from becoming a Bayesian in practice was my strong aversion to the Bayesian procedure for testing a precise null hypothesis against a composite alternative.

There are many statisticians today who believe that hypothesis testing in general and the testing of precise null hypotheses in particular are topics that have received attention out of all proportion to their importance. I am not one of them. Edwards, Lindman and Savage (1963), writing for an audience of psychologists and devoting more than half of their article to hypothesis testing, excused themselves as follows (p26):

"Our devotion of most of the rest of this paper to tests would be disproportionate, if we were not writing for an audience accustomed to think of statistics largely as testing."

As to the importance or otherwise of precise null hypotheses, Berger and Sellke (1987) adduce several reasons why they should often be used as convenient approximations for near-precise ones; but for me the crux of the argument on both points is that every time a decision is made to include or not to include an additional explanator in a regression, the choice is essentially one between a precise null hypothesis, namely that, "The parameter definitely has value zero," and its complementary alternative that, "The parameter's value
is not zero." There are few problems in statistics more pervasive than that one.

However, regardless of whether I am right or wrong concerning the intrinsic importance of this topic, it is unquestionably one where Frequentist and Bayesian inference can be and often are in serious conflict. It provides a suitable example for my thesis on that account alone, and I shall proceed with it on that basis.

It has long been understood that one of the most important differences between Frequentist and Bayesian inference is that, "the significance argument is based on the area under a curve and the Bayesian argument is based on the ordinate of the curve," (Lindley, 1957, p.190). This sounds harmless enough, but is almost responsible on its own for Bayesians regarding Frequentists as far too ready to reject a precise null hypothesis and for Frequentists regarding Bayesians as far too ready to accept one.

Consider first the Bayesians' attack on the Frequentist hypothesis test. Their arguments are particularly telling in the case where a single observation is taken from a uniform distribution. Suppose then that $Y$ is distributed uniformly between $(\mu - 0.5)$ and $(\mu + 0.5)$, that the null hypothesis is $H_0: \mu = 0$ and the alternative $H_1: \mu \neq 0$, and that the single observation is $y = 0.5 - \varepsilon$ where $\varepsilon$ is arbitrarily small. The relevant Frequentist statistic is the $p$-value: $\varepsilon$ for a one sided and $2\varepsilon$ for a two sided test. We will assume here that $\varepsilon$ is small enough for even the most exacting Frequentist to assert that there is overwhelming evidence against the null. For a Bayesian, however,
there is no shred of empirical evidence against $H_0: \mu = 0$, for the likelihood is uniform over the interval $-\epsilon$ to $1-\epsilon$ and zero elsewhere on the real line, so no value of $\mu$ is any more favoured by the evidence than $\mu = 0$.

If called upon to arbitrate between the two statisticians in this dispute, I would begin by saying that if the situation is really as they both describe it, then the Bayesian argument is compelling. I would, however, immediately qualify that judgement by noting that if $\mu$ is really zero, the premise that $Y$ is distributed uniformly between $(\mu - 0.5)$ and $(\mu + 0.5)$ must be highly questionable. A different alternative hypothesis, along the lines that $Y$ has mean zero and takes values exclusively between $-0.5$ and $+0.5$ but that there is a high probability of those values being close to $\pm 0.5$, would not be rejected by either line of reasoning. I conjecture, in fact, that whenever the two approaches give completely opposite answers, there is nearly always something wrong with the way the problem is specified.

This does not prevent the two approaches disagreeing quite substantially even when the problem is well specified. Suppose that instead of a single observation drawn from a uniform distribution we now have an observed mean, $\bar{y}$, of $n$ observations from $N(\mu, \sigma^2)$ where $\sigma^2$ is either known or sufficiently well estimated to be treated as known. Defining $t$ conventionally as $(\bar{y} - \mu)\sqrt{n} / \sigma$, a Frequentist carrying out a two-tailed test might describe the evidence against $H_0: \mu = 0$ as being "only mild" if $t = 1$ ($p = 0.32$), "significant" if $t = 2$ ($p = 0.046$), "highly significant" if $t = 3$ ($p < 0.003$) and "overwhelming" if $t = 4$ ($p < 0.0001$).
Bayesians, however, such as Edwards et al. (1963, pp44-45) and Berger and Sellke (1987, pp118-120), would argue that the four relevant likelihood ratios (i.e. the ratios of the normal density ordinates at $t = 1, 2, 3$ and 4 to the ordinate at $t = 0$) are 0.61, 0.14, 0.011 and 0.00034, and that these amount only to (almost?) no evidence, mild evidence, significant evidence and highly significant evidence respectively. Berger and Sellke add, "and even this may be overstating the evidence against $H_0$." The strength of the Bayesian argument on this point is that the likelihood ratios relate to the actual values observed, while the $p$-values relate predominantly to values that could have been observed but were not. I concede that ultimately this is a matter of taste, but the Bayesian arguments are the more convincing ones for me.

So far, we have concerned ourselves with Bayesian objections to the Frequentists' tendency to reject precise null hypotheses on insufficient grounds, but when we look at the difficulties that Bayesians have had in formulating such a test in a form capable of any public acceptance at all, we may well consider the stone throwers to be inhabiting a large and highly vulnerable glasshouse.

The normal distribution case will be quite sufficient to illustrate the problem. We may ask first, what is wrong with using the likelihood ratios mentioned two paragraphs above? Nothing, except that they tell only part of the story, and that the part least favourable to the null hypothesis. A genuine Bayesian test cannot accommodate an alternative hypothesis as vague as $H_1; \mu \neq 0$. A prior distribution for $\mu$ is required over the entire alternative parameter space. If that prior probability happens to be concentrated in a small region in the
vicinity of the observed mean, $\bar{y}$, then the empirical evidence favouring $H_0$ over $H_1$ is indeed the likelihood ratio evaluated at the observed value of $t$. If, however, the prior distribution over $\mu \neq 0$ is not concentrated in the "local region" of $\bar{y}$, then other likelihood ratios also start to enter the picture. These do not have the normal ordinate at $t = 0$ as their denominator, but smaller values, so the likelihood ratios are larger, and since these ratios provide the empirical evidence favouring $H_0$, that evidence (as summarized by the Bayes Factor) grows stronger and stronger as the prior density becomes more and more thinly spread over $\mu \neq 0$. Conversely, of course, the empirical evidence favouring $H_1$ becomes weaker and weaker.

Unfortunately, priors that are highly concentrated in any local region are by definition informative priors; that is, they provide information about the parameter, $\mu$, that is independent of the empirical evidence provided by the experiment currently being conducted. For legal and scientific purposes it is almost always convenient, and sometimes mandatory, to exhibit only the evidence provided by that experiment. This calls for an uninformative prior, and although for some probability distributions no prior can be uniquely uninformative, any prior with a claim to be considered as such will certainly be thinly spread over a large range of values of $\mu$. In the case of the normal distribution, a uniquely uninformative prior exists, namely the improper (or diffuse) uniform distribution over the entire real line.

This causes a real problem. When the prior density is diffuse, the Bayes Factor favouring $H_0$ is always infinite, so if any finite prior probability at all is associated with that hypothesis, then regardless of how many or how few
observations are taken, and no matter what values they take (as long as they are finite) the posterior probability associated with $H_0$ will be one. Nor can the problem be circumvented by attempting to define a prior distribution that is sufficiently close to being diffuse for practical purposes, for as the improper limit is approached, the Bayes Factor favouring $H_0$ becomes indefinitely large.

The (almost?) invariable Bayesian response to this dilemma has been to assert the need to resort to a subjective (and necessarily proper) specification of the alternative prior. This could be tolerable but for the fact that a wide range of permissible priors can yield virtually identical posterior distributions but wildly different Bayes Factors (Lavine and Wolpert 1993).

Edwards et al. claim (p4) that, "When a Bayesian procedure violates your intuition, reflection is likely to show the procedure to have been incorrectly applied," but despite the fact that the Bayesian test of a precise null hypothesis against a composite alternative has been producing counterintuitive results for nearly 40 years (since Bartlett 1957)—and arguably much longer, for the same problem is considered in Jeffreys (1948 and, it would seem, already in 1939). In this topic I have effectively taken Edwards et al. (1963) at their word. The reflection has required several years to take shape, but the outcome, described in Parts II and III of this Topic, seems to have been successful.

The lead-in to this discussion via Stone’s (1976) notorious Flatland problem also requires some explanation. Midway through my candidacy my then
supervisor, Prof R.L. Chambers, seeing which way my research was leading me, told me that if I was going to become a Bayesian I would need to come to terms with this conundrum. As even a brief perusal of the Discussion to Stone's paper will indicate, this was not a straightforward task. It did, however, provide a convenient example of both the strengths and the weaknesses of the Bayesian hypothesis testing procedure. It was strong for the "clockwise scenario" of Part I. The null and the alternative were there both precise, and an empirical Bayesian approach provided a solution different from and superior to the Classical (or Frequentist) version. It was, however, at the best questionable for the "anticlockwise scenario" of Part II, since the alternative hypothesis could be specified as diffuse, thereby guaranteeing \( H_0 \) a posterior probability of (almost) one. This served as a useful platform on which to demonstrate how that weakness could be circumvented by switching attention away from a reference Bayes Factor and towards a Reference Posterior Odds ratio. The properties of the that ratio itself are considered in Part III, and it is shown there that a strong link exists between it and the corresponding Frequentist test statistic.

One last point: Bayesian statisticians reading Parts I and II may find it unpalatable that the Bayesian "Barbara" should resort to the use of a Chi-squared statistic as a measure of goodness of fit. It is true that Bayesian statisticians do not at present avail themselves of this prerogative, but I see no reason why they should not do so. In its Frequentist form, the Chi-squared statistic is assumed to follow the central Chi-squared distribution, as prescribed by the null hypothesis that the goodness of fit is perfect. The one-sided \( p \)-value is then used to decide whether that null hypothesis needs to be
rejected or not: a decidedly un-Bayesian procedure. But there is nothing to prevent a Bayesian from postulating an alternative hypothesis in which the Chi-squared statistic follows a non-central Chi-squared distribution and where the non-centrality parameter has its own distribution over the positive half of the real line. True, there are no tables that can be looked up, but Bayesians are not usually frightened of number-crunching these days. My conjecture, for what it is worth, is that Bayesian statisticians have been passing up this opportunity because it involves testing a precise null hypothesis against a composite alternative, and because the subjective specification that they regard as necessary for the prior distribution for the non-centrality parameter is rather difficult to formulate.

3. THE NEED FOR CREATIVE TENSION

I have been accused of fence-sitting in my attitude to design-based and model-based sampling inference, and no doubt the same objection could be made to the way that I treat Bayesian and Frequentist inference. It has even been put to me that having one leg on each side of the fence must be decidedly uncomfortable, and in the long term likely to affect my prowess (as a statistician of course). I could perhaps answer in the same vein, "Not if I am of sufficient stature," but a more serious reply also seems called for.

I am not a fence-sitter. I am an eclectic, and where possible a synthesist. I am not just biding my time until I can see which way to jump; I am quite deliberately accepting the validity of both approaches and taking the best from each. This is particularly the case with design-based and model-based
sampling inference, for there I have indeed been able to construct a synthesis. The cosmetically calibrated estimator is fully valid and fully efficient under the design-based or randomization paradigm. It is fully valid (model-unbiased) under the model-based or prediction paradigm and, although it is not fully efficient, the extent of its inefficiency is seldom serious in practice (Topic A, Parts II and III).

I cannot make such a sweeping claim in relation to Bayesian and Frequentist inference, but I can nevertheless assert that now I am starting to feel reasonably at home in the Bayesian paradigm (and seeing it as more comprehensive and logically coherent than the Frequentist) I believe I can also see why it has failed to come to grips with two important problems, namely goodness of fit and the testing of a precise null hypothesis against a composite alternative. Further, I see this failure as having been brought about largely because Bayesians have regarded Frequentist inference as unworthy of attention. In both of these instances, Frequentist inference provides insights that can usefully be translated into Bayesian terms. In Topic B of this thesis, I will be providing the necessary translation for the hypothesis testing problem. I have already outlined in this introductory Chapter what might be achieved in the goodness of fit area, were the effort considered worthwhile.

I have paid no attention, however, to other forms of inference, such as the Fiducial and Likelihoodist approaches. This is not because I consider them unworthy of that attention, but purely because the Frequentist and Bayesian are the most popular and the ones with which I am most familiar. A widening of the synthetic approach to include other forms of inference might well reap
important dividends and, even if it were to prove fruitless, such an enterprise would still be desirable in principle.

Finally, where my critics see me as a fence-sitter, I tend to see them as blinkered or even one-eyed. What respectable physicist could assert, these days, that light should be treated exclusively either in terms of waves or of particles? The ability to hold apparently contradictory approaches in creative tension is also seen by Fowler (1981) as a sign of maturity in human development. [Fowler is one of many developmental psychologists that have modelled human development as taking place in discrete and recognisable stages. For a comparison of 16 such models see Wilber (c1980, especially pp180-183).] Fowler describes human development as taking place in seven stages, numbered from 0 to 6. He calls them "stages of faith," but is careful to say that this "faith" is not to be equated with religious belief. He succinctly distinguishes Stage 5 from Stage 4 by saying that it embraces "Many models," rather than only "One model."

I have taken some of the more specific distinctions from Fowler (1981, pp34-35) and displayed them in Table I. It will be seen that Fowler also regards openness to other groups of peoples' views as fairly central to the distinction between his Stages 4 and 5. [Wilber's (c1980, p181) equivalents to Fowler's
<table>
<thead>
<tr>
<th>Aspect</th>
<th>Stage 4</th>
<th>Stage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Role-taking</td>
<td>Mutual, with self-selected group or class—(Societal)</td>
<td>Mutual with groups, classes and traditions &quot;other&quot; than one's own.</td>
</tr>
<tr>
<td>Bounds of Social Awareness</td>
<td>Ideologically compatible communities with congruence to self-chosen norms and insights.</td>
<td>Extends beyond class norms and interests. Disciplined ideological vulnerability to &quot;truths&quot; and &quot;claims&quot; of outgroups and other traditions.</td>
</tr>
<tr>
<td>Locus of Authority</td>
<td>One's own judgment-as informed by a self-ratified ideological perspective. Authorities and norms must be congruent with this.</td>
<td>Dialectical joining of judgement-experience processes with reflective claims of others and of various expressions of cumulative human wisdom.</td>
</tr>
<tr>
<td>Form of World Coherence</td>
<td>Explicit system, conceptually mediated, clarity about boundaries and inner connections of system.</td>
<td>Multisystemic symbolic and conceptual mediation.</td>
</tr>
</tbody>
</table>

Stages 4 and 5 include Fromm's and Riesman's "Inner-directed Conformity" and their "Autonomy" respectively. In Fowler's Chapter 20, entitled "Stage 5. Conjunctive Faith," he likens the emergence of Stage 5 to being
"something like

"Realizing that the behavior of light requires that it be understood both as a wave phenomenon and as particles of energy.

"Discovering that the rational solution or 'explanation' of a problem that seemed so elegant is but a painted canvas covering an intricate, endlessly intriguing cavern of surprising depth.

"Looking at a field of flowers simultaneously through a microscope and a wide-angled lens."

While far from claiming that I am uniformly able to view the world from the perspective of Fowler's Stage 5, I can honestly say that I had been attempting to do so long before I came across Fowler, or even Wilber. This thesis is a continuation of that attempt. When about halfway through my investigations, I tried to explain what they were about to one of my non-statistical friends, who happened to be something of a philosopher. After listening for a few minutes he commented, "But this is a work of reconciliation!" He could not have accorded me a higher accolade.
REFERENCES

based Prediction and Design-Based Estimation Be Reconciled?"

Editions. Wiley. New York:

Edwards, W., Lindman, H. and Savage, L.J. (1963), "Bayesian Statistical
Inference for Psychological Research," Psychological Review, 70,
193-242. (Reprinted as Part I of Robustness of Bayesian Analyses, J.
references in the text have been taken from this reprint.)

Supplementary Information in Standard Sampling Procedures,

Fowler, J. (1981), Stages of Faith: The Psychology of Human Development

Godambe, V.P. (1955), "A Unified Theory of Sampling from Finite
Populations," Journal of the Royal Statistical Society, Ser. B., 17, 269-
278.

Hansen, M.H. and Hurwitz, W.N. (1943), "On the Theory of Sampling from

Horvitz, D.G. and Thompson, D.J. (1952), "A Generalization of Sampling
Without Replacement from a Finite Universe," Journal of the
American Statistical Association, 47, 663-685.

Oxford: Oxford University Press.


TOPIC A

Design-based and Model-based Inference in Business Survey Sampling

"All models are wrong but some are useful."

G.E.P. Box

Part I. Understanding the Differences

Early survey statisticians faced a puzzling choice between randomized sampling and purposive selection but, during the 1940s and early 1950s, Neyman's design-based or randomization paradigm became generally accepted as standard. It remained virtually unchallenged until the early 1970s when Royall and his co-authors produced an alternative paradigm based on statistical modelling. This revived the old idea of purposive selection under the new name of "balanced sampling." Most survey statisticians remained unconvinced by this, but they had long been making an implicit use of models in such techniques as ratio estimation and as time went on they started to use them more explicitly, though still working exclusively within the paradigm of design-based inference. This tendency eventually led to the respecification of the randomization paradigm in a revised version, now usually described as "model assisted survey sampling."

The fundamental disagreement between the supporters of this approach and the advocates of Royall's model-based paradigm still continues. It is shown here that both the difficulties inherent in the selection of a balanced sample and the role that computers are playing in overcoming them appear to have influenced the course of this debate quite substantially.

KEY WORDS: Balanced sampling; Model assisted survey sampling; Randomization; Representative Principle.
1. INTRODUCTION

Modern survey sampling if we date it from Kiaer (1897), is now almost a century old. For most of that time survey designers have had a choice between two options: randomized sampling or purposive selection. Indeed, for the first half of that period most survey designers were choosing between them on an arbitrary basis, apparently without serious fear of criticism. In the language used by Kuhn (1970), this was a preparadigmatic period. This was followed by some 25 years (between, say, 1945 and 1970) during which Neyman's (1934) design-based or randomization paradigm went virtually unchallenged. Finally, during the early 1970s the choice re-emerged; this time as one between two clearly articulated paradigms. Design-based sampling continued to dominate the field—of recent years in the modified version known as model assisted survey sampling—but it faced a serious challenge from the purely model-based approach, pioneered by Royall (1970, 1971) and consolidated by Royall and Herson (1973a & b), which used purposive selection under another name. At the time of writing, the issue between the two approaches remains unresolved. It will be demonstrated in Part I that each has its own strengths, that they tend to be complementary, and that there is something substantial to be gained by using them in combination.

2. ORDINARY DESIGN-BASED INFERENCE

The rigorous theory of design-based inference dates essentially from Neyman (1934), but it was definitively established as a dominant paradigm with the
textbooks written by Yates (1949) and Deming (1950), and even more influentially those written by Hansen, Hurwitz and Madow (1953), and Cochran (1953).

The design-based or randomization paradigm uses sample selection probabilities to provide the basis for its inferences. For the most part it is only the first and second order inclusion probabilities of the individual population units that are relevant. (An $r$th order inclusion probability is the probability of inclusion in sample of $r$ specified population units together.) These probabilities need to be known by the analyst, and it is a further requirement that all the first order inclusion probabilities be strictly positive. Unless these two conditions are met, no satisfactory point estimates can be made of the simple quantities that constitute the focus of an enumerative sample survey—population totals, population means and ratios of such totals or means. A sample that meets these two requirements is known as a probability sample. If, in addition, satisfactory estimates are required of the design-variances of these point estimates, the second order inclusion probabilities must also be knowable and strictly positive.

The properties of a design-based estimator are defined in terms of its behaviour over repeated sampling (i.e. in expectation over the set of all possible samples permitted by the sample design). In its pure form, design-based inference rests on what may be termed the Representative Principle. Imagine a population divided into three strata on the basis of size. From each of these strata a simple random sample of fixed size is drawn with equal probabilities without replacement. The probabilities of inclusion in sample are
unity for each (large) unit in the first stratum, 1/10 for each (medium-sized) unit in the second stratum and 1/100 for each (small) population unit in the third. A unit in the first stratum clearly needs only to represent itself, and the Representative Principle accordingly specifies that it have weight unity. But for every unit included in sample in the second stratum there are nine others in the population that are not in the sample, so each sample unit has weight ten. Similarly, each sample unit in the third stratum represents itself and 99 other population units not in the sample, so it has weight 100. More generally, for any probability sample, the Representative Principle requires that the weight given to each unit included in sample should be the reciprocal of its probability of inclusion. The properties of the estimator based exclusively on this principle were first considered in detail by Horvitz and Thompson (1952), and it is consequently known as the Horvitz-Thompson (HT) estimator.

The HT estimator may be specified as follows. Consider a population of \( N \) units from which a sample \( s \) is drawn, containing \( n \) units. The item of interest, denoted here by \( y \), is known as the survey variable. (Typically there are many survey variables, but they can usually be considered one at a time.) The \( N \)-vector of the population \( y \) values is denoted by \( y \) and the population total for \( y \) (the sum of the \( N \) \( y \)-values in \( y \)) by \( T(y) \). If the \( j \)th unit in the sample \( s \) has the \( y \) value \( y_j \) and its (first order) inclusion probability is \( \pi_j \), the HT estimator of \( T(y) \) is defined to be

\[
\hat{T}_{HT}(y) = \sum_{j \in s} y_j \pi_j^{-1}.
\]
In the special case where the $\pi_j$ are all equal, the HT estimator reduces to the ordinary expansion estimator $\hat{T}_{EXP}(y) = \left(\frac{N}{n}\right) \sum_{j=1}^{n} y_j$.

Much ingenuity has been used to elucidate various properties of the HT estimator, notably by Hanurav (1962, 1968). The most important ones are as follows:

(i) **Design-unbiasedness.** The HT estimator is uniquely unbiased over repeated sampling within the class of homogeneous linear estimators

$$\hat{T}(y) = \sum_{j=1}^{n} c_j y_j$$

where the coefficient $c_j$ depends only on the identity of the population labelled $j$ and not on the identities of the other population units included in sample or on any other feature of the sample design.

(ii) **Design-consistency.** As $N$ and $n$ approach infinity together in the manner described by Brewer (1979) [or in any of the variants on that manner suggested by Robinson and Tsui (1979), Isaki and Fuller (1982) and Särndal and Wright (1984)], the ratio of $\hat{T}_{HT}(y)$ to the actual population total $T(y)$ tends to unity; i.e. $\lim_{N,n \to \infty} (\hat{T}_{HT}(y)/T(y)) = 1$.

(iii) **Calibration on the inclusion probabilities.** If the sample size, $n$, is fixed, so that the sum of the inclusion probabilities is also $n$, the HT estimator estimates this sum without error; i.e.
This last property makes the HT estimator particularly suitable for estimating the totals of items whose values are closely proportional to the \( n_j \). However \( n_j \) may be the only variable against which the HT estimator is calibrated. If it is calibrated on any other variable, that variable is necessarily proportional to the \( n_j \).

The fact that the calibration property holds only for designs for which the sample size is fixed draws attention to an important feature in all design-based estimation, namely the need for the sample size to be at least approximately fixed in order for useful inferences to be made "over all possible samples". In particular, since design-variance is defined in such terms, both that variance itself and sample estimates of it can be very misleading where the observed sample is far from being "typical" in size.

An example may be useful here. Suppose a simple random sample is to be selected without replacement from a very large population. The sample size is determined by a preliminary draw of a random integer between 0 and 99. If that random integer is positive the sample size is to be 990, but if it is zero the sample size is to be only 10. Conditionally on the random integer being positive the design-variance is approximately \( S^2 / 990 \) (where \( S^2 \) is the population variance) but conditionally on the random integer being zero, the design-variance is approximately \( S^2 / 10 \). Both these conditional variances
are obviously meaningful. But design-based variances are nearly always defined unconditionally "over all possible samples". In this case the unconditional design-variance is approximately \( S^2 / 500 \). So with probability 0.99 the design-variance defined "over all possible samples" overestimates the meaningful (conditional) design-variance by a factor of nearly two, while with probability 0.01 it underestimates it by a factor of 50. [A similar point is made in Lahiri (1968, p4, para6).]

However, the condition that the sample size be at least approximately fixed is not \textit{sufficient} to guarantee meaningful design-based inferences. Even the stronger condition that all the possible samples must be of identical size can also be insufficient if that sample size is small. It can be seen from the HT estimator formula that the usefulness of this estimator really depends on the quite strong assumption that for every unit in sample having inclusion probability \( \pi_j \) there will be approximately \( \pi_j^{-1} - 1 \) units in the remainder of the population having somewhat similar characteristics. This second assumption is fundamental; it is what lies behind the Representative Principle itself. Very small probability samples are particularly unlikely to satisfy it. Basu's (1971) elephant fable provides an entertaining account of how an undue concentration on expediency in the sampling design can lead to such a breakdown, with devastating consequences.

Subject to this strong second assumption, however, (which actually subsumes the first one) the HT estimator is a useful estimator with several appealing properties. It is used in almost all instances of design-based estimation—
either in its original form or as a building block towards the construction of derivative estimators that can be tailor-made to fit particular survey situations.

It does, however, have one unfortunate drawback. This is the role played by the second order inclusion probabilities, $\pi_{jk}$, both in the formulae for its design-variance and in the sample estimators of that variance. If the sample size is fixed at $n = \sum_{\infty} \pi_j$ (an integer) the design-variance of the HT estimator is

$$V_p \hat{T}_{HT}(y) = \sum_{j=1}^{N} \sum_{k=1}^{N} (\pi_j \pi_k - \pi_{jk})(y_j \pi_j^{-1} - y_k \pi_k^{-1})^2$$

and its estimator,

$$\hat{V}_p \hat{T}_{HT}(y) = \sum_{j=\infty}^{\infty} \sum_{k=\infty}^{\infty} \pi_{jk}^{-1} (\pi_j \pi_k - \pi_{jk})(y_j \pi_j^{-1} - y_k \pi_k^{-1})^2,$$

is design-unbiased provided $\pi_{jk} > 0, \forall k < j$ (Sen 1953, Yates and Grundy 1953). However if $\pi_{jk} = 0$ for any $k < j$, the Sen-Yates-Grundy estimator is downwardly biased. Moreover, if any of these $\pi_{jk}$ is small positive, $\hat{V}_p \hat{T}_{HT}(y)$ has a large design variance and in consequence is described as "unstable" (Rao and Bayless 1969, Bayless and Rao 1970). If the method of sample selection chosen is one that is easy to use, the $\pi_{jk}$ can often take values that make the sample estimator of variance highly unstable or even biased. Conversely, if the method of sample selection is one yielding $\pi_{jk}$ values that guarantee $\hat{V}_p \hat{T}_{HT}(y)$ to be relatively stable, that method is
typically complicated to operate. Even then, the instability of $\hat{\sigma}_p^2$ is still markedly greater than that of the corresponding variance estimator for the Hansen-Hurwitz estimator of $T(y)$, which could have been used had the sample been selected with the same unequal probabilities but with replacement. Finally, regardless of the method used to select the sample with unequal probabilities without replacement, it is almost invariably tedious to calculate (or even to program the calculation of) the actual values of the $\pi_{jk}$.

[See Brewer and Hanif (1983, Chapter 3) for a fuller discussion of these points.]

Other design-based estimators, some of which bear little obvious resemblance to the HT estimator, have been suggested from time to time, notably by Des Raj (1956), by Murthy (1957) and by Rao, Hartley and Cochran (1962). The Des Raj and Rao-Hartley-Cochran estimators have manageable variance estimators, but they are themselves less directly based on the Representative Principle, the sample coefficients applied to $y_j$ depending on considerations other than the identity and properties of the $j$th sample unit. A small amount of efficiency is lost in consequence, and a certain amount of elegance is also surrendered. Since these alternative estimators have attracted little interest compared with that long commanded by the HT estimator and the estimators derived from it, and since it will be shown in Section 6 that the problems associated with the $\pi_{jk}$ can be overcome by using a combination of design-based and model-based inference, there seems little point in considering them further.
The simplest of the estimators derived from the HT estimator is the Horvitz-Thompson ratio estimator,

$$\hat{T}_{HT}(y) = \frac{\hat{T}_{HT}(y)}{\hat{T}_{HT}(x)} T(x),$$

(Brewer 1963, Hájek 1971). $\hat{T}_{HT}(y)$ itself is not design-unbiased, but it is asymptotically design-unbiased (Brewer 1979) and design-consistent. It is calibrated on the variable $x_j$ but not on the $\pi_j$. When the $\pi_j$ are equal, it reduces to the classical ratio estimator, $\hat{T}_r(y) = (\hat{T}_{EXP}(y) / \hat{T}_{EXP}(x)) T(x)$.

Another important derivative estimator is a generalization of the HT ratio estimator known as the Generalized Regression Estimator, a full consideration of which belongs in the next section.

3. MODEL ASSISTED SURVEY SAMPLING

The HT ratio estimator, $\hat{T}_{HT}(y)$ above, provides a good example of the way in which models of the population were long used in an implicit fashion by design-oriented survey statisticians. The essential difference between ordinary design-based inference and model assisted survey sampling is that the latter brings such implicit assumptions into the open and seeks explicitly to optimize a design-based estimator within a defined class, using as a working hypothesis that the population being sampled is structured exactly in accordance with an assumed model. Early examples of this approach may be found in Cochran (1953, 1963) and also in Brewer (1963) and Foreman and
Brewer (1971), but it was definitively established as the dominant version of the design-based paradigm with the publication of Särndal, Swensson and Wretman (1992).

If the HT estimator is regarded as the archetypal estimator of $T(y)$ under the ordinary design-based approach, its counterpart in model assisted survey sampling is the Generalised Regression Estimator of Cassel, Särndal and Wretman (1976):

$$
\hat{T}_{GREG}(y) = \hat{T}_{HT}(y) + [T(X) - \hat{T}_{HT}(X)]\hat{\beta}_{GREG}
$$

$$
= \sum_{j \in S} y_j \pi_j^{-1} + \left\{ T(X) - \sum_{j \in S} x_j' \pi_j^{-1} \right\} (W'X_v)^{-1}W'y_v,
$$

where $X$ is an $N \times p$ matrix of regressor variables, $X_v$ is the $n \times p$ matrix of its sample values, $y_v$ is the $n$-vector of sample values of $y$ and $W_v$ is any $n \times p$ weights matrix that ensures $\hat{\beta}_{GREG}$ to be a consistent estimator of $\beta$ in the explicitly assumed model for $y$, namely

$$
\xi: \quad y_j = x_j' \beta + \varepsilon_j; \quad E_\xi \varepsilon_j = 0, \quad E_\xi \varepsilon_j^2 = \sigma^2 \alpha_j^2, \quad E_\xi \varepsilon_j \varepsilon_k = 0, \forall k \neq j
$$

or more succinctly,

$$
\xi: \quad y = X\beta + \varepsilon; \quad E_\xi \varepsilon = 0, \quad E_\xi \varepsilon \varepsilon' = \sigma^2 A^2,
$$
where $A = \text{diag}(a_j)$. In Särndal et al (1992) the preferred expression for $\hat{\beta}_{\text{GREG}}$ is

$$
\hat{\beta}_{\text{GREG}} = (X_i' A_i^{-2} \Pi_i^{-1} X_i)^{-1} X_i' A_i^{-2} \Pi_i^{-1} y_i,
$$

where $\Pi = \text{diag}(\pi_j)$ and $A_i, \Pi_i$ are the $n \times n$ matrices of the sample values of $A_i, \Pi_i$. This expression for $\hat{\beta}_{\text{GREG}}$ implies that $W_i' = X_i' A_i^{-2} \Pi_i^{-1}$, but since the second term in the formula for $\hat{T}_{\text{GREG}}(y)$ is a small adjustment term of order $n^{-1}$, the value of $\hat{T}_{\text{GREG}}(y)$ is robust against changes in $W_i$ when $n$ is large.

The important properties of $\hat{T}_{\text{GREG}}(y)$ are that it is design-consistent and asymptotically design-unbiased, and that it is also calibrated against all the variables in the columns of $X$. It is not, however, calibrated against the $\pi_j$ unless the vector of the $\pi_j$ is already itself in the column space of $X$.

Moreover, the HT ratio estimator, $\hat{T}_{\text{HTR}}(y)$, can be regarded as the special case of $\hat{T}_{\text{GREG}}(y)$ when there is only a single column in $X$ and $a_j^2 \propto x_j$. If in addition $W_i = 1_n$, it reduces still further to the ordinary or classical ratio estimator

$$
\hat{T}_r(y) = (\hat{T}_{\text{EXP}}(y) / \hat{T}_{\text{EXP}}(x))T(x) = \left( \sum_{j \in x} y_j / \sum_{j \in x} x_j \right) T(x).
$$

4. MODEL-BASED INFERENCE

All the estimators so far considered have been essentially design-based; that is to say, the probability structure on which the statistical inferences drawn from them are based is the set of the probabilities, $p$, with which the various
possible samples, \( s \in S \), could be selected. [The probability of selecting sample \( s \) is conventionally denoted by \( p(s) \).] Where the estimator uses the HT estimator as a building block—as is usually the case—substantive inferences are based on the \( \pi_j \) and inferences concerning the reliability of those substantive inferences involve both the \( \pi_j \) and the \( \pi_{ik} \).

For samples that are not probability samples, a radically different type of inference is necessary. Such inferences were regularly made from purposively selected samples before the arguments for the exclusive use of probability samples became generally accepted (in the late 1940s), but these inferences were essentially ad hoc in nature and their shortcomings were incisively exposed by Neyman (1934). The serious advocacy for the use of an alternative source of inference commenced with Royall (1970).

Royall argued that survey sampling was out of step with statistics as a whole. Statisticians working in other fields used their data to build models and analysed them in those terms (using model-based inference) but survey statisticians had allowed themselves to be seduced into using an entirely irrelevant source of probability structure, not related to the data themselves but only to the manner in which they had been collected. He suggested that in many instances a suitable model for inferential purposes was one that will here be written as

\[
\xi_j: \quad \gamma_j = \beta x_j + u_j; \quad E_{\xi_j} u_j = 0, \quad E_{\xi_j} u_j^2 = \sigma^2 \alpha_j^2, \quad E_{\xi_j} u_j u_k = 0, \forall k \neq j.
\]
Here the operator $E_{\xi_{1}}$ denotes the expectation over all realizations of the model $\xi_{1}$, the $a_{j}^{2}$ of the variance function are assumed to be known, and $\beta$ and $\sigma^{2}$ are parameters that can be estimated from the data Model $\xi_{1}$ is a special case of the model $\xi$ defined in Section 3.

(For many purposes Royall also made the further assumption that the variance function was proportional to $x_{j}$. This assumption seems to have been made partly for mathematical convenience, partly because it was often close to the truth, and partly because his most important inferences were reasonably robust against likely departures from this specification.)

Royall first showed that the best linear unbiased predictor (BLUP) of $T(y)$ under $\xi_{1}$ was

$$\hat{T}_{BLUP}(y) = \sum_{j \in s} y_{j} + (T(x) - \sum_{j \in s} x_{j}) \hat{\beta}_{BLUE},$$

where $\hat{\beta}_{BLUE}$, the best linear unbiased estimator of $\beta$, is given by

$$\hat{\beta}_{BLUE} = \sum_{j \in s} y_{j}x_{j}a_{j}^{-2} / \sum_{j \in s} x_{j}^{2}a_{j}^{-2}.$$

The model-based variance of $\hat{\beta}_{BLUE}$ is

$$V_{\xi}(\hat{\beta}_{BLUE}) = \frac{\sigma^{2}}{\sum_{j \in s} x_{j}^{2}a_{j}^{-2}}.$$
so that of $\hat{T}_{BLUP}(y)$ is

$$V_\xi[\hat{T}_{BLUP}(y)] = \frac{(T(x) - \sum_{j \in S} x_j)^2 \sigma^2}{\sum_{j \in S} x_j^2 \sigma_j^2}. $$

Since in normal circumstances $a_i^2/x_i^2$ is a decreasing function of $x_i$, $V_\xi(\hat{\beta}_{BLUE})$ and $V_\xi[\hat{T}_{BLUP}(y)]$ take their smallest possible values when the sample consists of the $n$ units having the largest values of $x_j$. This is then the optimal sample under model $\xi^1$, and Royall initially suggested that such samples should be used routinely in survey sampling.

Royall's further assumption, that the variance function was proportional to $x_j$, ensured that $\hat{\beta}_{BLUE}$ simplified down to $\hat{\beta}_R = \sum_{j \in S} y_j / \sum_{j \in S} x_j$ and hence ensured also that $\hat{T}_{BLUP}(y)$ was identical with the classical ratio estimator,

$$\hat{T}_R(y) = T(x) \left\{ \sum_{j \in S} y_j / \sum_{j \in S} x_j \right\}.$$

This retention of a well-known estimator combined with a sample consisting of the $n$ units having the largest values of $x_j$ appeared to offer quite remarkable gains in efficiency and involve comparatively little in the way of practical complications. It was soon found, however, that estimates based on this "optimal" sample were highly sensitive to departures from the model $\xi^1$. Small units had only to behave slightly differently from large units for the bias in $\hat{T}_{BLUP}(y)$ to become non-negligible.
Royall and Herson (1973a) responded to this problem by considering the most plausible type of model breakdown (which they took to be $y_j$ having a general polynomial dependence on $x_j$ rather than a homogeneous linear one) and showing that the selection of a balanced sample—essentially one that was a miniature of the original population—rather than a simple random one, would provide the necessary protection against model breakdown. Specifically, they showed that if the true model generating the population was

$$
\xi[\delta_0, \delta_1, \ldots, \delta_L; v(x)]: \quad y_j = \delta_0 \beta_0 + \delta_1 \beta_1 x_j + \cdots + \delta_L \beta_L x_j^L + \epsilon_j[v(x_j)]^{0.5}
$$

$$
\epsilon_j \sim \text{i.i.d.}(0, \sigma^2)
$$

(where the $\delta_i$ take the values unity or zero, so that $\xi_i$ above corresponds to $\xi[0, 1; x]$) then the model-bias of $\hat{T}_R(y)$ could be reduced to zero by ensuring that the sample means of the $x_j^i$ were equal to the corresponding population means for all values of $i$ for which $\delta_i = 1$. They defined a balanced sample of order $L$ as one for which the sample mean, $\bar{x}_s^{(i)}$, of the $x_j^i$ was equal to its population mean, $x^{(i)}$, for $i = 1, 2, \ldots, L$, and denoted such a sample by $s(L)$.

They then showed that for a sample that is at least $s(1)$ the ratio estimator, $\hat{T}_R(y)$, reduced to the simple expansion estimator, $(N / n) \sum_{j \in s} y_j$, which is also the BLUP of $T(y)$ under $\xi[1; 1]$; also more generally that if the sample is $s(L)$ the BLUP under an $L$th polynomial regression model would also reduce to the expansion estimator, provided that the variance function was

*When discussing or quoting from papers by Royall and his co-authors, some changes have been made to their notation in order to conform to that used elsewhere in this article. In particular, the subscripts $j$ and $l$ have been used in place of $k$ and $j$ respectively.*
proportional to $x_i^l$ for some $l = 0, 1, \ldots, L$ and that the regression function itself contained the $l$th degree term $\beta_l x_i^l$.

The other principal advantages attendant on the use of a balanced sample were summarized by the authors as follows:

"When $\xi[0,1:x]$ is the true model, the ratio estimator is optimal for any $s$. If the ratio estimator is used, but the model is actually $\xi[1,1:v(x)]$, a bias is incurred. The possibility of such a bias can be avoided by the choice of a balanced sample $s(1)$ instead of the optimal sample [the one composed of the $n$ units whose $x$-values are largest]. Protection against a certain type of error in the model $\xi[0,1:x]$ is gained, and some efficiency under this model is lost. If the additional conditions $x_{i(s)} = x_{i(s^*)}$, $l = 2, 3, \ldots, L$, are also satisfied, then the unbiasedness of the ratio estimator is insured under any model of the form $\xi[\beta_0, \beta_1, \ldots, \beta_L : v(x)]$, yet no additional loss in efficiency is incurred. After balance on the means is achieved, balance on the higher moments provides protection against many types of error in the model $\xi[0,1:x]$ at no cost in terms of additional loss of efficiency under $\xi[0,1:x]$.”

The relationship of balanced sampling to simple random sampling was seen to be of considerable importance. Simple random sampling usually produces
samples that are approximately balanced. Indeed, a simple random sample of
given size will yield a sample with any required degree of approximate balance
with a calculable probability. Conceding this, Royall and Herson proceeded
to comment as follows:

"However, simple random sampling produces samples which
are not all good approximations to \( s(L) \). If the actual
regression function is a polynomial but not a straight line
through the origin and if, as happens with predictable
frequency, the randomly selected sample is badly unbalanced,
then the ratio estimator is biased. In such a case the fact that
simple random sampling will produce balanced samples on
average is of little consolation. For providing protection
against bias in the ratio estimator when a polynomial
regression model applies, deliberate balancing is preferable to
unrestricted random selection."

Towards the end of this paper, however, Royall and Herson gave their
qualified approval to the use of restricted randomization as a useful step
towards the achievement of an approximately balanced sample:

"[R]andom selection will at least justify some degree of
confidence that the selected sample is approximately
representative. Thus what is needed is assured balance on at
least the first moment of \( x \), and reasonable certainty of
adequate balance on various other (perhaps unidentified)
variables. One way to accomplish this is through some form of restricted randomization, say random selection from among all samples $s$ which are approximately balanced with respect to $x$.

In the second part of their paper, Royall and Herson (1973b) showed that if balanced sampling were to be applied within size strata and used with the stratum-by-stratum (or separate) ratio estimator, the result would be even greater robustness against model breakdown. Moreover, if the generating model was in fact $\xi[0,1:x]$, much of the efficiency that had been lost by using a simply balanced sample as opposed to the "optimal" sample consisting of the largest $n$ units would be recovered. Given a cost function constant from stratum to stratum, the optimal allocation of sample units to strata was closely analogous with Neyman allocation; with $n_h/N_h$, the sampling fraction for stratum $h$, being proportional to $\overline{x}_h^{0.5}$ under $\xi[0,1:x]$ and to $[v(\overline{x}_h)]^{0.5}$ under $\xi[\delta_0,\delta_1,...,\delta_t,...,\delta_l;v(x)]$.

The authors also recommended that stratum boundaries be drawn in such a manner that the optimal allocation of sample units to strata would ensure the number of sample units selected to be constant from stratum to stratum. This was to facilitate the selection of a suitably balanced sample within each stratum. Heuristic arguments and some limited empirical evidence both suggested that these two objectives could both be achieved under $\xi[0,1:x]$ if the values of $N_h^2\overline{x}_h^{(1)}$ were themselves held roughly constant from stratum to stratum.
The whole notion of balancing was, in fact, a return to the ideal that had been aimed for back in the preparadigmatic days by those survey statisticians who practiced "purposive selection". Thus Gini and Galvani (1929) wrote that the aim of purposive selection was, "... to obtain a sample which would be representative of the whole country with respect to its chief demographic, social, economic and geographic characteristics," and the Commission set up by the International Statistical Institute to study the Representative Method (1926) described such samples as being, "made up by purposive selection of groups of units which it is presumed will give the sample the same characteristics as the whole."*

It may be conjectured in hindsight that if Gini and Galvani had had access to computers comparable in size to the ones used by Royall and Cumberland (1981a,b), the history of sample survey design and inference might have been entirely different. Neyman (1934, p582 footnote) quotes them as saying, "nobody could under-appreciate the difficulty in a stratification of the communes simultaneously with regard to different characters."† No wonder, for there were 8,354 of them. To make their task manageable they used the 214 districts (circondari) as their sample units instead, selecting a mere 29 of them. Had they been able to use communes they would have had a much larger number of sample points on which to base their

* The quotations from Gini and Galvani (1921, p1) and from the Commission set up by the International Statistical Institute to study the Representative Method (1926, p376) are taken from Neyman (1934, pp559-560). The first of these is presumably Neyman's own translation. The original reads, "... ottenere un campione rappresentativo di tutti il Paese, rispetto alla sue principali caratteristiche demografiche, sociali, economiche e geografiche, ..."

† The original reads, "non è chi non veda quali ostacoli si incontrerebbero nell'eseguire tale «stratificazione» simultaneamente rispetto a diversi caratteri."
estimates, and that would have helped to make their estimates much more reliable and probably much more accurate.

Royall and Cumberland (1981a,b) selected their approximately balanced samples using restricted randomization. Successive randomly drawn samples were rejected until one was found that was balanced to a preassigned level of accuracy. The mean squared errors of $\hat{T}_r(y)$ obtained using these approximately balanced samples were smaller and distinctly more stable than those found for the set of unrestricted random samples. The model-based variances were also more meaningful than the design-based variance, in that each model-based variance referred to a particular sample (the one actually selected) whereas the design-based variance was by definition based on the potential variability from sample to sample over the feasible set, $s \in S$, and could therefore make no distinction between those (well-balanced) samples that accurately reflected the characteristics of the population and the ill-balanced ones that did not.*

In Parts II and III which follow, the potentials of both model assisted survey sampling and the fully model-based paradigm are compared in for stratified sampling (Part II) and for unequal probability sampling (Part III). Part IV summarizes the results obtained in the first three Parts and assesses the extent to which the desired reconciliation of the two approaches can be regarded as successfully achieved.

*Although the detailed consideration of variance estimation is outside the scope of this thesis, it may be worth mentioning here Royall and Eberhardt's (1975) use of the model-based approach for this purpose. This paper highlighted the fact that a frequently used estimator of design-based variance suffered from a paradoxical property; namely that the lower the model-based variance for a given sample—and hence the more accurate the sample estimate—the higher the estimate of the design-based variance. Since the accuracy of the sample estimate would be judged by the estimate of its design-based variance, it followed that the more accurate it was in fact, the less accurate it would be judged to be. However, an alternative design-based estimator free from this paradox is also available.
II. Stratified Balanced and Stratified Random Sampling

Suppose that the sampling strategy to be used for a particular survey must involve both a stratified sampling design and the classical ratio estimator, but that, within a stratum, a choice is allowed between simple balanced sampling and simple random sampling; then which should the survey statistician choose? The balanced sampling strategy appears preferable, both in terms of efficiency under a frequently used working model and of robustness against that model breaking down, but the randomized design has certain countervailing advantages. These include the simplicity of the selection process and an established public acceptance that randomization is "fair." It transpires that nearly all the advantages of both schemes can be secured if simple random samples are selected within each stratum, but a GLS regression estimator (with a diagonal variance matrix) is used instead of the classical ratio estimator. The regression estimator ensures calibration of the survey estimates on all the explanatory variables used, which is the same property that balanced sampling uses to ensure the robustness of the ratio estimator. It is less efficient than the ratio estimator (and unacceptably so if the sample size within a stratum is not a substantial multiple of the number of explanatory variables) but it is nevertheless efficient asymptotically and more flexible in its application than balanced sampling. Moreover, the random selection process is considerably easier to handle than the balanced one, particularly if permanent random numbers are being used to control sample rotation and/or overlap. Finally, the balanced sampling strategy restricts the statistician to the use of model-based inference whereas the randomized sample permits the use of design-based and model-based inference simultaneously.

KEY WORDS: Coverage Probability; Empirical Bayes; Noninformative prior.
1. WHERE BALANCED SAMPLING HAS THE EDGE

When comparing the advantages and disadvantages of using design-based and model-based sampling inference in the context of a stratified sampling design, it is helpful to note that the two approaches can easily result in the selection of readily comparable samples and the application of identical estimators. We have seen in Part I that the model-based approach assuming model \( \xi[0,1;x] \) naturally leads to the use of stratified balanced sampling and the stratum by stratum (or separate) ratio estimator,

\[
\hat{T}_{str}(y) = \sum_h \hat{T}(y_h) = \sum_h T(x_h) \left\{ \frac{\sum_j y_{hj}}{\sum_j x_{hj}} \right\},
\]

with the allocation of the sample among strata such that the sampling fraction, \( n_h/N_h \), is proportional to \( \bar{x}_h^{0.5} \) and with the stratum boundaries drawn so that the \( N_h^2 \bar{x}_h^{(1)} \) are roughly constant over \( h \).

Under the traditional design-based approach of stratified random sampling there would be only three differences worth mentioning:

1.1 The stratum boundaries

In practice, these would be determined using whatever appeared to be convenient, such as "less than $100,000, $100,000-199,000, $200,000-499,999, $500,000-999,999 and $1,000,000 and over"). However, since the optimum specification of stratum boundaries is quite flat, and since experienced survey statisticians of all persuasions
usually manage to avoid serious departures from it, this difference can be ignored. We will suppose here that the two designs use the same stratum boundaries.

1.2 The type of ratio estimator

If the samples within the individual strata were small, the across-stratum or combined ratio estimator,

$$\hat{T}_{asR}(y_h) = \left[ \sum_h \left( \frac{N_h}{n_h} \sum_{j \in S_h} y_{hj} \right) \sum_h T(x_h) \right] \left/ \left[ \sum_h \left( \frac{N_h}{n_h} \sum_{j \in S_h} x_{hj} \right) \right] \right,$$

might be chosen in place of $\hat{T}_{str}(y)$. This would be to avoid "ratio-estimation bias", which is well known to decrease as $n^{-1}$ (Cochran 1963 p161). Further, under the working model $\xi[0,1:x]$, its model-expected value is zero, so except where some of the $n_h$ are very small, say fewer than six, there is little or no need to avoid it.

There is actually a paradox here. From the design-oriented approach, no model is postulated as generating the data, so it is assumed both that there will be a real difference in the average ratio of $T(y_h)$ to $T(x_h)$ from stratum to stratum and that there will be a design-bias in the ratio estimator. If the individual sample sizes, $n_h$, are large for each $h$, $\hat{T}_{str}(y)$ is preferred on the grounds of efficiency. If, however, one or more of the $n_h$ is small, then $\hat{T}_{asR}(y_h)$ is preferred, in order to increase the effective sample size and thereby reduce the bias. This can be interpreted as purchasing robustness against variability in the ratio of $T(y_h)$ to $T(x_h)$ at the possible expense of efficiency in estimating the overall ratio of $T(y)$ to $T(x)$ should that variability be negligible. However, from the model-oriented approach, if there is some confidence
that the actual population is capable of being modelled by $\xi[0,1:x]$, $\hat{T}_{aR}(y_h)$ is preferred on the grounds of efficiency, but if there is some doubt as to the suitability of $\xi[0,1:x]$ as a model, $\hat{T}_{uRAT}(y)$ is used to achieve robustness against model breakdown.

So from the design-oriented standpoint $\hat{T}_{nR}(y)$ is efficient and $\hat{T}_{aR}(y_h)$ is robust, while from the model-oriented standpoint the reverse is true! The paradox is easily explained, however, as soon as we ask under what assumptions each estimator is efficient and each is robust. $\hat{T}_{nR}(y)$ is efficient if the ratio of $T(y_h)$ to $T(x_h)$ varies over $h$, while $\hat{T}_{aR}(y_h)$ is efficient if the model parameter $\beta$ is constant over $h$.

Conversely, $\hat{T}_{nR}(y)$ is specifically robust against variability from stratum to stratum in a meaningful parameter $\beta_h$ while $\hat{T}_{aR}(y_h)$ is robust when the overall ratio of $T(y)$ to $T(x)$ is a poor measure of the relationship between $y$ and $x$, either in the whole or in any part of the population.

In any case, the difference between the two estimators is usually small and, if none of the $n_h$ is less than six, can reasonably be ignored. We will assume here that this is the case and that the stratum-by-stratum ratio estimator, $\hat{T}_{nR}(y)$, is used for both designs.

1.3 The type of sampling

For the model-based strategy the samples within each stratum would be selected in a balanced fashion. For the design-based strategy, simple random sampling would be used without replacement. In contrast to the situations described in 1.1 and 1.2 above, this difference is quite an important one. Balanced sampling is demonstrably superior to random sampling, both in terms of robustness and of efficiency.
1.3.1 Robustness. If the generating model is not $\xi[0,1; x]$ but $\xi[\delta_0, \delta_1, \ldots, \delta_L; v(x)]$, the estimator $\hat{T}_{sr}(y)$ will in general be model-biased, but not if the sample is $s(L)$. More generally, if the generating model is the $\xi$ of Part I, Section 3, namely

$$\xi: \quad y = X\beta + \varepsilon; \quad E_\xi \varepsilon = 0, \quad E_\xi \varepsilon\varepsilon' = \sigma^2 A^2,$$

where there are $p$ explanatory variables in the columns of $X$, then once again that estimator will in general be model-biased, but not if the sample is balanced on those variables. (This is true whether or not the particular explanatory variable, $x$, that appears in the definition of $\hat{T}_{sr}(y)$, is included or excluded from that balancing.)

Balancing can therefore make a very important contribution to the robustness of that estimator, over and above the "balancing in expectation" supplied by randomization. As pointed out by Royall and Herson (1973a, p887) "... if, as happens with predictable frequency, the randomly selected sample is badly unbalanced, then the ratio estimator is [model-]biased. In such a case the fact that simple random sampling will produce balanced samples on the average is of little consolation."

Ideally, the sample would be balanced on all $p$ explanatory variables, in which case $\hat{T}_{sr}(y)$ would actually be unbiased under the general model, $\xi$ above, but as Royall and Herson themselves say, there are practical problems in the selection of balanced samples, so:
"[W]hat is often needed is assured balance on at least the first moment of \( x \)
and reasonable certainty of adequate balance on various other (perhaps
unidentified) variables." (1973a p 888.)

Fortunately, however, even if a sample is only explicitly balanced on a single variable,
such as the first moment of \( x \), it will usually be better balanced even on the other
variables than a randomized sample of the same size. Indeed, the least favourable
situation for any given variable (in the double expectation over all possible samples
and all realizations of any relevant model) is for the balanced sample to be no worse
than the randomized one (Tam and Chan 1984).

1.3.2 Efficiency. Using \( \bar{x} \), \( \bar{x}_s \), and \( \bar{x}_n \) to denote the population, sample and non-
sample means respectively, Royall and Herson (1973a, p 887) reported on their
relevant findings as follows:

"The balanced sampling plan is again superior to simple random sampling if
bias is ignored and the two plans are compared in terms of the error variance
of the ratio estimator. This is easily established. When \( v(x) = x \) [as in our
case] the variance of \( \hat{T}_h(y) - T(y) \) is \( \sigma^2 \{ N(N-n)/n \bar{x}_s / \bar{x}_n \}. \) The average
value of this quantity over all possible samples of size \( n \) is
\( \sigma^2 \{ N(N-n)/n \bar{x}(Nc\bar{x} - n)/(N-n), \) where \( c \) is the average value of \( 1/\bar{x}_s \)
over all samples. Now a well-known inequality states that \( c \) is greater than or
equal to \( 1/\bar{x} \) with equality only in case all \( x \)'s are equal. Thus the average
variance, over all samples of size \( n \), is greater than \( \sigma^2 \{ N(N-n)/n \bar{x}, \) the
variance when the sample is balanced. This same result, that for the ratio
estimator the average error-variance over all samples of size \( n \) is greater than
the error-variance for the balanced sample, also applies when \( \nu(x) = 1 \), as can easily be shown."

Although stratified balanced sampling is slightly more efficient \textit{on average over all possible samples}, a badly balanced sample is about as likely to produce a substantially below average model-variance as one that is substantially above average. Moreover, since the difference in efficiencies depends entirely on the extent to which the average value of \( 1 / \bar{x}_s \) exceeds \( 1 / \bar{x} \), the most substantial edge of balanced sampling over random sampling is likely to be in the lowest size stratum, and not to be all that appreciable even there for samples of, say, six or more units. Nevertheless, it may be regarded as important to avoid the risk of selecting the kind of badly unbalanced sample that yields a high model-variance for \( \hat{T}_r(y) \). Once this has happened, then as with robustness, "the fact that simple random sampling will produce balanced samples \textit{on the average} is of little consolation."

2. WHERE RANDOMIZED SAMPLING HAS THE EDGE

So far, we have seen that when stratified balanced sampling and stratified random sampling are compared, the balanced sampling design provides better protection against extreme situations that can produce either high model-variance (under \( \xi[0,1:x] \)) or high model mean squared error (under the \( \xi \) of Part I, Section 3) and also tends to provide a little extra efficiency. It is now necessary to consider whether stratified random sampling has any compensating advantages. There are at least three such advantages worth mentioning.
2.1 Consistency in the implicit assumptions

Since the stratified random sample design is not tied to the assumption that the variance function is proportional to \( x_j \), it invites estimation of the individual stratum population variances—defined in terms of the variability in \((y_j - \beta x_j)^2\), in order to be relevant for \( \hat{T}_k(y) \). It therefore facilitates a more optimal allocation of the sample units among the strata than the proposal made by Royall and Herson (1973b) does. However, there is nothing to stop the stratified balanced sample design being "patched up" to match the stratified random one in this regard. The model would then be \( \xi[0,1: \nu(x)] \) rather than \( \xi[0,1:x] \) and there would be a slight logical inconsistency because \( \hat{T}_k(y) \) would still be used within each stratum, although it would no longer be the BLUP of \( T(y) \). It would, however, continue to be model-unbiased.

2.2 Simplicity in the selection process.

Most people of normal intelligence can quickly be trained to select a simple random sample, but it takes a computer to select a balanced one, especially if it is to be balanced on more than a single variable. We have seen in Part I, Section 4, how difficult Gini and Galvani (1929) found it to select a sample balanced on seven variables. Something of this advantage remains even in these days of high-speed computers. It is much simpler to accept the first random sample that comes along than it is to test it for balance, reject it if it is unsatisfactory, and repeat the process until a sample with satisfactory balance has been reached. Hardware, software, programming effort and computer time are all required, and each of these can be costly. Moreover, the Permanent Random Number technique (Ohlsson, 1995), now frequently being used to control sample rotation and response burden, is essentially
oriented to the selection of random samples. While it may be possible to program the
selection of balanced samples so as to control the outcomes in a similar fashion, such
a project presents a considerably more formidable task than that posed by Permanent
Random Number sampling.

2.3 Public acceptance.

There is an established perception, even among non-statisticians, that randomized
sampling is fair. This is unlikely to go away in the short or medium term—or, indeed,
as long as lotteries remain popular. There are, no doubt, ways in which balanced
sampling could be made as automatic (and in that sense as fair) as randomization, but
it may still be difficult to present it as fair. If, for instance, atypical units were
included in sample less frequently than typical ones, the selection procedure might
reasonably be perceived as inequitable. However unjust it may be to the highly
elegant concept of balanced sampling, the maxim that "justice must not only be done
but be seen to be done," tends to work against it.

3. CALIBRATION: GETTING THE BEST OF BOTH WORLDS

The advantages of simplicity and public acceptance both seem substantial
enough to prompt the question as to whether the stratified random design
could not itself be "patched up" to match or nearly match the stratified
balanced one in terms of efficiency and robustness. The answer is that a near
match is possible using calibration on all the columns of \( \mathbf{X} \). Such a
calibration can be achieved by using \( \hat{T}_{GREG}(y) \), instead of \( \hat{T}_{sr}(y) \). \( \hat{T}_{GREG}(y) \)
is the Generalized Regression Estimator of Part I, Section 3; however in this
instance, the $\pi_j$ within stratum $h$ are each equal to $n_h/N_h$, so $\hat{T}_{\text{GREG}}(y)$ reduces to a somewhat non-standard form of the stratum by stratum regression estimator, namely

$$\hat{T}_{\text{REG}}(y) = \sum_h \left[ \left( \frac{N_h}{n_h} \right) \mathbf{1}_{n_h} y_{hs} + \left( \frac{N_h}{n_h} \right) Y_{h} \mathbf{1}_{n_h} \mathbf{X}_{hs} \mathbf{X}_{hs}^{-1} \mathbf{Z}_{hs} \mathbf{X}_{hs}^{-1} \left\{ \mathbf{X}_{hs}^\prime \mathbf{Z}_{hs}^{-1} \mathbf{y}_{hs} \right\} \right],$$

where $\mathbf{Z}_{hs} = \mathbf{diag}(z_{hs})$ and $z_{hs}$ is any linear combination of the columns of $\mathbf{X}_{hs}$.

Reasons are given in Part III for suggesting that the most appropriate choice for $z_{hs}$ would be an overall measure of size for the $j$th sample unit.

Using $\hat{T}_{\text{REG}}(y)$ with a stratified random sample, rather than $\hat{T}_{\text{REG}}(y)$ with a sample balanced stratum by stratum on all the columns of $\mathbf{X}$, is analogous to using poststratification in preference to stratification. If the important axes of stratification are known, a stratified sample based on them should be preferred to a poststratified design; but if they are not known exactly, and an incorrect choice is made, the damage is done. Poststratification estimation is less efficient than stratification, but it is more flexible. Similarly, a sample balanced on all the important explanatory variables is better than a calibrated random sample. This is because for every additional regression variable that is not strictly necessary, there is an increase in the variability of the sample weights and a consequent increase in model-variance. However, if any genuine explanatory variable is omitted from the balancing process, $\hat{T}_{\text{REG}}(y)$ is irretrievably model-biased. Regression estimation operating on a stratified random sample will generally be less efficient than ratio estimation operating on a stratified balanced sample of the same size, but it is more flexible.
The extent of this inefficiency increases as the number of explanatory variables, \( p \), increases. As a very general guide, \( p / n \) should be small compared with one and \( p \) itself should not exceed a certain fixed limit. A convenient measure of the relative inefficiency of the randomization strategy is the effective sample size. Within stratum \( h \), the effective sample size is

\[
N_h^2 / \sum_{j \in s_h} w_{hj} (w_{hj} - 1),
\]

which takes its maximum possible value of \( n_h N_h / (N_h - n_h) \) when \( w_{hj} = N_h / n_h \) for all sample values of \( j \). If the introduction of an additional explanatory variable brings about a substantial reduction in the effective sample size, its usefulness must be questioned.

For any given set of explanatory variables, however, regression estimation is a good deal easier to execute than simultaneous balancing. From this point of view, a decision to use calibration rather than balancing could be seen as a preference for a simple sampling procedure combined with a relatively complex estimator over a simple estimator combined with a quite demanding sampling procedure.

There is one other potential advantage in the calibration approach that may well be of crucial importance. If \( \hat{T}_{sREG} (y) \), or for that matter \( \hat{T}_{aREG} (y) \), is used with a stratified random sample, it becomes possible to use model-based as well as design-based inference using the same estimator on the same set of data. For a full set of reasons as to why it is useful to be able to invoke both approaches simultaneously, it will be necessary to take into account how they eliminate the problem with the \( \pi_h \) under
unequal probability sampling (see Part III), but it is already worth making the point at this stage that model-based inference relates specifically to a particular sample whereas design-based inference relates to the entire set of possible samples that can be selected given that sample. Consequently, model-based inference is more appropriate for use at the analysis stage whereas, when designing any future probability sample, the likely variability over all possible samples that could be selected is the relevant issue, and it is then that design-based inference comes into its own.

4. SUMMARY

If a stratified balanced sampling strategy and a stratified random sampling strategy are both optimized for stratum by stratum ratio estimation, the two become very similar and differ importantly only in the inescapable fact that in the one case the samples are balanced within strata and that in the other case they are randomized.

The balanced sampling strategy is the more robust of the two, because randomization cannot guarantee a good balance in every case, and because for a badly balanced sample the ratio estimator can be substantially model-biased—in terms of the working model $\xi[0,1:v(x)]$—and may also have a substantially higher model-variance. The balanced sampling strategy is also more stable in its model-variance and, on the average over all possible samples, slightly more efficient.
The only substantial countervailing advantages of the randomization strategy are the current state of its public acceptance, its greater simplicity in the selection process and the possibility it affords of tapping in to the sample control afforded by Permanent Random Number selection. However if generalized regression estimation is substituted for ratio estimation in the randomized strategy, the latter can be improved both in robustness and efficiency to a point where it becomes comparable with the balanced one, and it also acquires the further substantial advantage of greater flexibility in the choice of explanatory variables.

The number of explanatory variables used in the equation should not be too large, however, or the inefficiency of the randomization strategy vis-à-vis the balancing strategy will become excessive. The effective sample size formula provided in Section 3 can be used as a convenient measure of this inefficiency.
III. Unequal Probability Sampling

The results of Part II are here extended to cover samples selected using unequal inclusion probabilities. The results obtained are heuristically instructive and on occasion the practical advantages can be substantial. Since the design-variance cannot be estimated without knowing the second order inclusion probabilities, the anticipated variance is estimated instead. This is the appropriate variance to consider for future sample designs but, for evaluating the sample actually surveyed, model-variance is more to the point. Two quite different techniques (cosmetic and calibration estimation) are used to produce equivalent expressions that are interpretable both as design-based regression estimators and model-based predictors. A large sample estimator is given for the variance of the model-based predictor.

An unanticipated spin-off is a simple and effective method for eliminating unacceptably small sample weights. This is put to the test using empirical farm data. In some extreme circumstances it is logically impossible to eliminate all such weights and at the same time retain all the desired calibration constraints, but the problem virtually disappears if the inclusion probabilities do not increase too rapidly with size. Cosmetic calibration is then comparable in accuracy with ordinary regression estimation, even after all unacceptably small weights have been eliminated. It also compares favourably with ridge regression and is only marginally less efficient than Nadaraya-Watson nonparametric regression.

KEY WORDS: Cosmetic calibration; Negative sample weights; Unequal probability sampling.
1. INTRODUCTION

It is not difficult to modify the results of Part II to cover samples selected using unequal probabilities, but a word as to why this might be considered useful will be in order first. There are both advantages and disadvantages in replacing a stratified random sample by an unequal probability selection scheme across the whole population.

Advantages

(i) Under $\xi[0,1:\nu(x)]$ the optimum inclusion probabilities are proportional to $\nu(x_j)^{0.5}$ across the entire population. ["Optimum" in this context refers to the minimizing of the anticipated variance (Isaki and Fuller 1982) of $\hat{T}_{GREG}(y)$. Since $\hat{T}_{GREG}(y)$ is model-unbiased, its anticipated variance is both the expectation of $V_{\xi}\hat{T}_{GREG}(y)$ over all realizations of $\xi$ and the expectation of $V_{\xi}\hat{T}_{GREG}(y)$ over repeated sampling.]

(ii) Elimination of size stratification permits a finer stratification on axes other than size.

(iii) Circumstances occasionally arise where unequal probability sampling is the natural way to proceed, regardless of any formal consideration of optimality.

Disadvantages

(iv) Elimination of size stratification involves a loss of robustness against the possibility of a nonlinear dependence of $y$ on $X$. 
(v) The design-variance and its estimators involve the second order inclusion probabilities, $\pi_{jk}$, and hence the associated difficulties described in Part I, Section 2.

(vi) The actual process of selection is made more complicated.

Advantage (i) produces only a modest gain in efficiency under any particular specification of $\xi[0,1:v(x)]$, but if the parameter $\beta$ varies appreciably from type to type (e.g. between pairs of closely related industries), Advantage (ii) can lead to substantial reductions in anticipated variance. An example of a survey where Advantage (iii) comes into play is the Vegetable Chemical Use Survey currently being conducted by the National Agricultural Statistical Service of the US Department of Agriculture. In this survey, a farm merits inclusion in sample to the extent that it produces important quantities of specified crops. A farm that merits inclusion on account of more than one crop is accorded the largest of the inclusion probabilities merited by each crop separately.

Disadvantage (iv) can be substantial if the relationship between $y$ and $X$ is strongly nonlinear, but if the nonlinearity itself can be modelled, $X$ can be expanded accordingly and linearity restored. The difficulties contingent on Disadvantage (v) can be circumvented if anticipated variance is estimated instead of design-variance.

Disadvantage (vi) can be minimized by careful choice of selection method. The simplest form of Permanent Random Number sampling that could be used is Poisson sampling (Hájek 1964, 1981), but the sample size is a random number and subject to appreciable variance. Collocated sampling (Brewer, Early and Joyce 1972, Brewer, Early and Hanif 1984) is a reasonably simple modification of Poisson sampling that
reduces that variance quite considerably. Synchronized Sampling is the form of Permanent Random Number sampling used by the Australian Bureau of Statistics (Hinde and Young 1984). It was devised for stratified random sampling, in which context it provides a fixed sample size for all practical purposes. The exceptions occur only as a result of "births" and "deaths", and even then reasonably seldom. Unfortunately its generalization to unequal probability sampling, though quite possibly achievable given sufficient effort, is not straightforward. Unless and until this generalization is achieved, collocated sampling would seem to provide the best option.

In summary, there is always some limited gain in accuracy to be obtained from using unequal probability sampling, on some occasions they can be substantial, and there are no offsetting disadvantages that cannot be overcome with a reasonable amount of additional effort. Although in most instances the simplicity offered by a stratified random design will undoubtedly be the overriding consideration, there will also be times when unequal probability sampling is clearly preferable. (It is in fact being used in the Vegetable Chemical Use Survey mentioned earlier in this Section.)

If it is also taken into account that some properties of combined inference can readily be appreciated only when the more general case is considered explicitly, the argument for taking the trouble to describe it becomes reasonably compelling. As a preliminary, we consider in some detail the nature of an important concept we have already mentioned several times in this Section, namely anticipated variance.
2. DESIGN-VARIANCE AND ANTICIPATED VARIANCE

It was noted in Part I, Section 2, that the design-variance of the HT estimator was

\[ V_p \hat{T}_{HT}(y) = \sum_{j=1}^{N} \sum_{k=1}^{N} (\pi_j \pi_k - \pi_{jk})(y_j \pi_j^{-1} - y_k \pi_k^{-1})^2. \]

If the generating model is the \( \xi \) of Part I, Section 3, i.e.

\[ \xi : \quad y_j = x_j' \beta + \epsilon_j, \quad E_{\xi} \epsilon_j = 0, \quad E_{\xi} \epsilon_j^2 = \sigma^2 \alpha_j^2, \quad E_{\xi} \epsilon_j \epsilon_k = 0, \quad \forall k \neq j \]

and if the estimator used is the Generalised Regression Estimator,

\[ \hat{T}_{GREG}(y) = \hat{T}_{HT}(y) + \{T(X) - \hat{T}_{HT}(X)\} \hat{\beta}_{GREG}, \]

then the design-variance of \( \hat{T}_{GREG}(y) \) is given asymptotically for large \( n \) by

\[ V_p \hat{T}_{GREG}(y) = \sum_{j=1}^{N} \sum_{k=1}^{N} (\pi_j \pi_k - \pi_{jk}) \{y_j - x_j' \beta \} \pi_j^{-1} \{y_k - x_k' \beta \} \pi_k^{-1} \]
\[ = \sum_{j=1}^{N} \sum_{k=1}^{N} (\pi_j \pi_k - \pi_{jk}) \{\epsilon_j \pi_j^{-1} - \epsilon_k \pi_k^{-1}\}^2. \]

The anticipated variance (Isaki and Fuller 1982) of \( \hat{T}_{GREG}(y) \) is the expectation of \( V_p \hat{T}_{GREG}(y) \) over all realizations of \( \xi \). This can be expressed asymptotically for large \( n \) as follows:
This expression was first derived by Godambe (1955) and is the minimum possible anticipated variance for any design-unbiased estimator of $T(y)$.

Since $\hat{\beta}_{GREG}$ is always to be a consistent estimator for $\beta$, and since also

$$E_{\xi} (y_j - x'_j \hat{\beta})^2 (\pi_j^{-1} - 1) = \sigma^2 a_j^2 (\pi_j^{-1} - 1),$$

the anticipated variance of $\hat{T}_{GREG}(y)$ (for large $n$) can be estimated by

$$\sum_{j \in r} (y_j - x'_j \hat{\beta}_{GREG})^2 (\pi_j^{-1} - 1)$$.
3. THE COSMETIC APPROACH TO COMBINED ESTIMATION

The anticipated variance is an appropriate variance to estimate for sample design purposes, but for the analysis of any particular sample the ordinary model-based variance is a more logical choice. To estimate this it is convenient to have $\hat{T}_{\text{GREG}}(y)$ expressed in predictor form.

There are two ways in which this can be done, either by using Cosmetic Estimation (Särndal and Wright 1984) or by using Calibration Estimation (Deville and Särndal 1992). The Cosmetic approach requires immediately that there be an estimator of $\beta$, (here written $\hat{\beta}_{\text{cos}}$) such that the generalized regression and the predictor forms are numerically equal, i.e.

$$\hat{T}_{\text{cos}}(y) = \Pi_1 y + (1'\Pi_1^{-1}X - 1'\Pi_1^{-1}X_0)\hat{\beta}_{\text{cos}} = 1'y + (1'\Pi_1^{-1}X - 1'\Pi_1^{-1}X_0)\hat{\beta}_{\text{cos}}.$$

where $\Pi_1$ is the $n \times n$ diagonal matrix of the sample $\pi_j$.

Although the right hand expression above is now already in predictor form, that predictor cannot generally be the BLUP, as that would require $\hat{\beta}_{\text{cos}}$ to be $\hat{\beta}_{\text{BLUE}}$. This, as will be shown directly, would generally lead to a contradiction.

The two expressions for $\hat{T}_{\text{cos}}(y)$ immediately above, when taken together, imply that $1'(\Pi_1^{-1} - I)(y_0 - X_0\hat{\beta}_{\text{cos}}) = 0$. To meet this condition we first impose the weak requirement that $\hat{\beta}_{\text{cos}}$ be of the projection form.
\[ \hat{\beta}_{\text{cos}} = (Q'X_s)^{-1}Q'y, \] where \( Q_s \) is any \( n \times m \) matrix of rank \( m \) such that \( Q'^TX_s \) is positive definite. We then also impose the stronger requirement that \( (\Pi^{-1}_s - I_n)I_n \) be spanned by the column space of \( Q_s \), i.e. that there must be some \( m \)-vector \( \alpha_s \) such that \( Q_s \alpha_s = (\Pi^{-1}_s - I_n)I_n \). Then,

\[ 1_s' (\Pi^{-1}_s - I_n) (y_s - X_s \hat{\beta}_{\text{cos}}) = \alpha_s' Q_s' [y_s - X_s (Q'X_s)^{-1} Q'y_s] \]
\[ = \alpha_s' [Q' - Q'X_s (Q'X_s)^{-1} Q']y, \]
\[ = 0, \]

satisfying the condition. If, however, \( \hat{\beta}_{\text{blue}} \) is used to estimate \( \beta \), then \( Q_s = A^{-2}_s X_s \) and the column space of \( A^{-2}_s X_s \) does not, in general, span \( (\Pi^{-1}_s - I_n)I_n \). Since \( (\Pi^{-1}_s - I_n)I_n \) must be in the column space of \( Q_s \), either \( X \) or \( A \) must be changed. The possibility of modifying \( X \) was explored in Brewer (1995), but subsequent empirical tests (some of which are described in Section 6) have indicated that it is more appropriate to change \( A \).

To this end, we replace \( A^{-2} \) in the expression \( \hat{\beta}_{\text{blue}} = [X' A^{-2} X_s]^{-1} X' A^{-2} y \), by \( Z_s^{-1} (\Pi^{-1}_s - I_s) \) where \( Z_s \) is an \( n \times n \) diagonal matrix, \( Z_s I_n \) being any arbitrary linear combination of the columns of \( X_s \). We then have that

\[ \hat{\beta}_{\text{cos}} = [X' Z_s^{-1} (\Pi^{-1}_s - I_s) X_s]^{-1} X' Z_s^{-1} (\Pi^{-1}_s - I_s) y_s , \]
\[ Q_s = (\Pi^{-1}_s - I_n) Z_s^{-1} X_s \],
\[ \alpha_s = (Z_s^{-1} X_s)^{-1} I_n \] and \( Q_s \alpha_s = (\Pi^{-1}_s - I_n)I_n \) as required. The Cosmetic Estimator of \( T(y) \) can then be written
\[ \hat{T}_{\cos}(y) = I_n' \Pi_s^{-1} y_s + (I_n' X - I_n' \Pi_s^{-1} X_s) \hat{\beta}_{\cos} \]
\[ = I_n' \Pi_s^{-1} y_s + (I_n' X - I_n' \Pi_s^{-1} X_s)[X' (Z_s^{-1} (\Pi_s^{-1} - I_n) X_s)]^{-1} X_s' Z_s^{-1} (\Pi_s^{-1} - I_n) y_s. \]

An informed guess can be made as to an optimal choice for \( Z_s \). If the non-zero elements \( z_j^{-1} (\pi_j^{-1} - 1) \) in the diagonal matrix \( Z_s^{-1} (\Pi_s^{-1} - I_n) \) were proportional to the corresponding non-zero elements in \( A^{-2} \), namely \( a_j^{-2} \), then \( \hat{\beta}_{\cos} \) would indeed be identical with \( \hat{\beta}_{BLUE} \) and \( \hat{T}_{\cos}(y) \) would be the BLUP of \( T(y) \) under the general model \( \xi \) that was re-introduced in Section 2. It therefore makes sense to aim for the \( z_j^{-1} (\pi_j^{-1} - 1) \) to be as closely proportional to the \( a_j^{-2} \) as possible.

Suppose we make the following three strong assumptions:

(i) that the \( \pi_j \) are optimally determined, and hence proportional to the \( a_j \),
(ii) that the \( a_j \) are proportional to a particular linear combination of the columns of \( X \), denoted here by \( \tilde{z} \), and
(iii) that each one of the \( \pi_j \) is small compared with unity;

then the choice \( z = \tilde{z} \) ensures that the \( z_j^{-1} (\pi_j^{-1} - 1) \) are closely proportional to the \( a_j^{-2} \), and is therefore nearly optimal. The likelihoods and consequences of breakdowns in these three assumptions are considered below.

(i) In practice, Assumption (i) is unlikely to break down badly. Most survey statisticians are well aware of the fact that optimum allocation requires the
sample fraction in a given stratum to be approximately proportional to the stratum average of the $a_j$, and allocate their samples accordingly.

(ii) We know that $z$ must be a linear combination of the columns of $X$, so if there is any such linear combination on which the $a_j$ depend, even if not strictly proportionately, it will be difficult to find a better choice. Hence a breakdown in Assumption (ii) is unlikely to change the situation substantially.

(iii) Serious breakdowns in Assumption (iii), however, will occur more often than not. This would cast doubt on the usefulness of the choice $z = \tilde{z}$, except for the fact that there is another issue at stake here. The factor $(\pi_j^{-1} - 1)$ (rather than simply $\pi_j^{-1}$) appears in the formula for $\hat{\beta}_{cos}$ precisely because this estimator is not designed to apply to the population as a whole, but only to the non-sample portion of it.

This is best appreciated by first considering an extreme case. If a particular population unit is to be included in sample with certainty, the Representative Principle of Part I, Section 2, requires that it should make no contribution whatsoever to the estimate for the non-sample portion of that population. An important part of the price paid for the kind of robustness that the Representative Principle can sometimes offer—that is, when the conditions discussed at the end of Section 2 actually hold—is that the model information supplied by the $j$th sample unit must be diluted by the factor $(1 - \pi_j)$ before being applied to an estimate of the balance of the population. Hence the appropriateness of the choice of $z = \tilde{z}$ is not affected by any breakdown in Assumption (iii), even by a severe one.
In summary, even in situations where Assumption (i) does not hold exactly and the other two do not hold at all, the natural choice for $z$ is still whatever linear combination of the columns of $X$ most nearly resembles the variance function.

4. THE CALIBRATION APPROACH TO COMBINED ESTIMATION

Identical formulae for estimating $T(y)$ and $\beta$ can be obtained following the Calibration procedure described by Deville and Särndal (1992). The essence of this approach is that the sample weights should be as "close" as possible to the $\pi_j^{-1}$, subject to the condition that, for every variable in the columns of $X$, the sample estimate defined by these weights should be without error. The "closeness" is defined by an arbitrary distance function, but for our present purposes the appropriate function is

$$D = (w_i - \sigma_i)'Z_i(\Pi_i^{-1} - I_n)(w_i - \sigma_i) + 2\lambda'(X_i'1_n - X_i'w_i),$$

where $w_i$ is the $n$-vector of the sample weights $w_i$, $\sigma_i = \Pi_i^{-1}1_n$ is the $n$-vector of the inverse inclusion probabilities $\pi_j^{-1}$ and $\lambda'$ is a $1 \times p$ row vector of undetermined multipliers. Differentiating with respect to $w_i$,

$$\frac{\partial D}{\partial w_i} = 2Z_i(\Pi_i^{-1} - I_n)(w_i - \sigma_i) - 2X_i\lambda.$$ 

Solving $\frac{\partial D}{\partial w_i} = 0$ yields
\[ \mathbf{w}_t = \mathbf{\omega}_t + (\Pi_t^{-1} - \mathbf{I}_n)^{-1} \mathbf{Z}_t^{-1} \mathbf{X}_t \lambda, \]

\[ \mathbf{X}_t' \mathbf{w}_t = \mathbf{X}_t' \mathbf{\omega}_t + \mathbf{X}_t' (\Pi_t^{-1} - \mathbf{I}_n)^{-1} \mathbf{Z}_t^{-1} \mathbf{X}_t \lambda, \]

\[ \lambda = [\mathbf{X}_t' (\Pi_t^{-1} - \mathbf{I}_n)^{-1} \mathbf{Z}_t^{-1} \mathbf{X}_t]^{-1} \mathbf{X}_t' (\mathbf{w}_t - \mathbf{\omega}_t) \]

\[ = [\mathbf{X}_t' (\Pi_t^{-1} - \mathbf{I}_n)^{-1} \mathbf{Z}_t^{-1} \mathbf{X}_t]^{-1} [\mathbf{T}(\mathbf{X}) - \mathbf{X}_t' \mathbf{\omega}_t] \text{ and} \]

\[ \mathbf{w}_t = \mathbf{\omega}_t + (\Pi_t^{-1} - \mathbf{I}_n)^{-1} \mathbf{Z}_t^{-1} \mathbf{X}_t [\mathbf{X}_t' (\Pi_t^{-1} - \mathbf{I}_n)^{-1} \mathbf{Z}_t^{-1} \mathbf{X}_t]^{-1} [\mathbf{T}(\mathbf{X}) - \mathbf{X}_t' \mathbf{\omega}_t]. \]

The corresponding Calibration Estimator is

\[ \hat{T}_{\text{CAL}}(y) = \mathbf{\omega}_t' \mathbf{y}_t + [\mathbf{T}(\mathbf{X}) - \mathbf{X}_t' \mathbf{\omega}_t] [\mathbf{X}_t' (\Pi_t^{-1} - \mathbf{I}_n)^{-1} \mathbf{Z}_t^{-1} \mathbf{X}_t]^{-1} \mathbf{X}_t' (\Pi_t^{-1} - \mathbf{I}_n)^{-1} \mathbf{Z}_t^{-1} \mathbf{y}_t \]

\[ = \mathbf{1}_n' \Pi_t^{-1} \mathbf{y}_t + (\mathbf{1}_n' \mathbf{X} - \mathbf{1}_n' \Pi_t^{-1} \mathbf{X}_t) \hat{\beta}_{\text{COS}}. \]

This is exactly the same formula as was obtained for \( \hat{T}_{\text{CO}}(y) \) in Section 3.

**5. COMPARING THE COSMETIC AND CALIBRATION APPROACHES**

Since the two approaches yield identical estimators for both \( \beta \) and \( T(y) \), these will be described from now on as Cosmetic Calibration Estimators and be denoted by \( \hat{\beta}_{\text{COSCAL}} \) and \( \hat{T}_{\text{COSCAL}}(y) \) respectively. \( \hat{T}_{\text{COSCAL}}(y) \) can be written in five different forms, with at least four different nuances of interpretation.
(i) The Cosmetic Generalized Regression Estimator

\[ \hat{T}_{\text{coscal}}(y) = 1_n' \Pi_n^{-1} y + (1_n' X - 1_n' \Pi_n^{-1} X_s) [X'_s Z_s^{-1} (\Pi_s^{-1} - I_n) X_s]^{-1} X'_s Z_s^{-1} (\Pi_s^{-1} - I_n) y_s. \]

(ii) The Cosmetic Predictor

\[ \hat{T}_{\text{coscal}}(y) = 1'_n y_s + (1'_n X - 1'_n X_s) [X'_s Z_s^{-1} (\Pi_s^{-1} - I_n) X_s]^{-1} X'_s Z_s^{-1} (\Pi_s^{-1} - I_n) y_s. \]

(iii) The Calibration Estimator (short form)

\[ \hat{T}_{\text{coscal}}(y) = w'_n y_s. \]

(iv) The Calibration Estimator (long form)

\[ \hat{T}_{\text{coscal}}(y) = w'_n y_s + (1'_n X - w'_n X_s) [X'_s Z_s^{-1} (W_s - I_s) X_s]^{-1} X'_s Z_s^{-1} (W_s - I_s) y_s, \]

where \( W_s \) is the \( n \times n \) diagonal matrix of the \( w_j \), i.e. \( W_s 1_n = w_s. \) [The second term of (iv) can be added to the short form shown in (iii) because \( 1'_n X - w'_n X_s = 0. \)]

(v) The Calibration Predictor

\[ \hat{T}_{\text{coscal}}(y) = 1'_n y_s + (1'_n X - 1'_n X_s) [X'_s Z_s^{-1} (W_s - I_s) X_s]^{-1} X'_s Z_s^{-1} (W_s - I_s) y_s. \]
This form can be derived from (v) using exactly the same steps as were used in Section 4 to show that (i) and (ii) were equivalent. However, the second term of (v) is not zero. It is in fact the leading term.

Further, because the predictors (ii) and (v) are equivalent, there are two forms for $\hat{\beta}_{\text{coscal}}$. These are,

(vi) The Cosmetic Form

$$\hat{\beta}_{\text{coscal}} = [X^tZ^{-1}(\Pi^{-1} - I_n)X]^{-1}X^tZ^{-1}(\Pi^{-1} - I_n)y,$$

and

(vii) The Calibration Form

$$\hat{\beta}_{\text{coscal}} = [X^tZ^{-1}(W_s - I_n)X]^{-1}X^tZ^{-1}(W_s - I_n)y.$$

6. THE MODEL–VARIANCE OF $\hat{T}_{\text{coscal}}(y)$ AND ITS ESTIMATION

The detailed consideration of model-variance and its estimation are outside the scope of this thesis, but it will be shown here that there is a simple and plausible estimator of the model-variance of $\hat{T}_{\text{coscal}}(y)$ that can be used when the sample size is large. (The estimation of terms other than the leading term can be performed in several ways and it is not yet obvious which should be recommended.)

The model-variance of $\hat{T}_{\text{coscal}}(y)$ is, by definition,
\[ E_\varepsilon [\hat{T}_{\text{COSCAL}}(y) - T(y)]^2 = E_\varepsilon \left[ \sum_{j \in S} w_j y_j - \sum_{j=1}^N y_j \right]^2 \]

\[ = E_\varepsilon \left[ \sum_{j \in S} (w_j - 1) y_j - \sum_{j \in S} y_j \right]^2 \]

\[ = \sigma^2 \left[ \sum_{j \in S} (w_j - 1)^2 a_j^2 + \sum_{j \in S} a_j^2 \right] \]

\[ = \sigma^2 \left[ \sum_{j \in S} w_j (w_j - 2) a_j^2 + \sum_{j=1}^N a_j^2 \right]. \]

Since it may not be easy to estimate \( \sum_{j=1}^N a_j^2 \) (other than by a HT estimator) it is worthwhile exploring whether the model-variance of \( \hat{T}_{\text{COSCAL}}(y) \) can be usefully approximated. The expression above is less than \( \sigma^2 \sum_{j \in S} w_j a_j^2 \) but greater than \( \sigma^2 \sum_{j \in S} w_j (w_j - 2) a_j^2 \) and the approximation \( \sigma^2 \sum_{j \in S} w_j (w_j - 1) a_j^2 \) is attractive on several grounds. These include the fact that it takes the value zero when all the \( w_j \) are equal to unity, and its resemblance to the estimator of the Anticipated Variance suggested at the end of Section 2, which may now be written \( \sum_{j \in S} \pi_j (\pi_j - 1)(y_j - x'_j \hat{\beta}_{\text{COSCAL}})^2 \).

Since for large sample sizes \( \hat{\beta}_{\text{COSCAL}} \) is close to \( \beta \), a natural estimator to use for the model-variance of \( \hat{T}_{\text{COSCAL}}(y) \) is \( \hat{V}[\hat{T}_{\text{COSCAL}}(y)] = \left( \sum_{j \in S} w_j (w_j - 1)(y_j - x'_j \hat{\beta}_{\text{COSCAL}}) \right)^2 \).

Note, however, that both \( \hat{\beta}_{\text{COSCAL}} \) and \( \hat{V}[\hat{T}_{\text{COSCAL}}(y)] \) can underestimate their corresponding actual variances substantially if the sample size is small. The leading term is of order \( N^2 n^{-1} \) and the largest neglected term is of order \( N^2 n^{-2} \). A more
79

detailed study of variance estimation in the context of combined estimation has yet to
be undertaken.

7. THE PROBLEM OF UNACCEPTABLY SMALL SAMPLE
   WEIGHTS

It was pointed out in Part I, Section 2, that a strong condition had to be
fulfilled before the Representative Principle underlying design-based
inference could be regarded as useful. This was that for every sample unit
included with probability \( \pi_j \) there should be approximately \( \pi_j^{-1} - 1 \) units with
reasonably similar properties in the non-sample portion of the population. In
some instances there are good reasons for this assumption; for instance where
the sample is a large one and the inclusion probabilities are an explicit
function of size, as in size-stratified sampling.

The manner in which the Cosmetic Calibration weights, \( w_k \), are constructed,
however, implies that they are better indexes of the relevant properties than the
\( \pi_k^{-1} \) are themselves. In fact, had the inclusion probabilities used been \( w_k^{-1} \),
and had the sample then selected been identical with the one actually selected,
that sample would have been perfectly balanced on all the variables in \( X \).
(Note, however, that the sample would have been balanced on the intercept
term only if the column space of \( X \) was already spanning \( 1_N \).)

So there is a sense in which the inverse weight, \( w_k^{-1} \), can be thought of as
analogous to an inclusion probability. Sample units with large weights (and
hence small \( w_k^{-1} \)) can be considered as typical in their characteristics in that
they "represent" large numbers of population units. Sample units with smaller weights are still to some extent typical, but they "represent" fewer population units. A sample unit with weight unity is only on the borderline of being typical. It does not represent any other unit. A sample unit with a weight less than one is definitely atypical. It does not even represent itself. Worse still, a sample unit with negative weight is quite perversely unrepresentative. For a small enough domain, it can actually produce negative estimates of total. Its presence in the sample is a "rare event". Yet it must be part of the population, or it would not be in the sample.

The obvious procedure to adopt for a unit with \( w_j < 1 \) is to delete it both from the sample and the sample frame, recalculate the \( w_j \) so that the remaining sample units estimate only the remaining population units, and then add the deleted unit on at the end as an atypical extra. This, of course, is precisely what many design-oriented survey statisticians have been doing with "outlying observations" for forty years or more. It can now be seen as the natural thing to do to accommodate rare and influential observations as well.

8. AN EMPIRICAL STUDY USING FARM DATA

The actual performance of cosmetic calibration as compared with certain alternative estimation procedures has been studied using data obtained from two farm surveys conducted by the Australian Bureau of Agricultural and Resource Economics.
Chambers (1996) carried out a comparison of various estimation strategies using economic and production data collected from a sample of 904 farms in the annual Australian Agricultural and Grazing Industries Survey (AAGIS) and Australian Dairy Industry Survey (ADIS) in the late 1980s. Since the original selection was made using stratified random sampling and the sample fractions were higher for the larger farms, the data set contains a disproportionate number of large farms as compared with Australian farms generally. Moreover, since beef farms tend to be large and dairy farms small, the former are over-represented and the latter under-represented. The complete data set contains a relatively large subset (farms with wheat) and a small one (farms with dairy cattle), both of which follow a model of the type $\xi$, reasonably well, together with two other large subsets (farms with sheep and farms with beef cattle) that are not at all well modelled by $\xi$, the nonlinearity in the subset of farms with beef cattle being particularly striking. These properties make it a useful and exacting data set for testing purposes.

The survey variables of interest here are four components of income (those from Wheat, Beef, Sheep and Dairy sales) and Total Income. Each component has a natural supplementary or "benchmark" variable that is its most important factor of production. These are Hectares of Wheat, Numbers of Beef Cattle, Numbers of Sheep and Numbers of Dairy Cattle respectively. The supplementary variable used for Total Income is Dry Sheep Equivalents (DSE) defined as $[(12 \times \text{hectares of Wheat}) + (8 \times \text{number of Beef Cattle}) + (12 \times \text{number of Dairy Cattle}) + \text{number of Sheep}]$. 
For his own purposes, Chambers generated three sets of stratified random (sub)samples from the 904 farm "population". Each set consisted of 500 samples of 100 farms. Those in Set 1 were simple random samples selected without replacement. Sets 2 and 3 were stratified random samples. For Set 2, Chambers used a "Compromise" allocation that was about midway between simple random sampling and the "Optimal" allocation that he used for Set 3. (This optimality was defined in terms of proportioning the sample fractions to the standard deviations of the DSE variable, and corresponded roughly to sampling with inclusion probabilities proportional to "size" as measured by the DSE.) The four sampling fractions used for Set 2 were 50/665, 25/166, 18/52 and 7/12. For Set 3 they were 30/665, 29/166, 29/52 and unity (12/12).

For the present study, three additional sets were selected, each again consisting of 500 samples of 100 farms. The inclusion probabilities used in each set were proportional to a fractional power of the DSE. For Set 4 that power was 0.60, for Set 5 it was 0.75, and for Set 6 it was 0.90. In each case there was also a completely enumerated sector. It was smallest for Set 4 and largest for Set 6.

Chambers calculated sample weights and root mean squared errors for each of Sets 1-3 using six different estimators. The first of these, "RATIO", was the HT ratio estimator based on each survey variable's natural supplementary variable. For his stratified random samples, this estimator was also the across-stratum or combined ratio estimator. He calculated this only as a basis of comparison for other estimators, holding it to be essentially unsatisfactory
in that the sample weights differed from one survey variable to another. In his Section 1, Motivation and Summary, he commented on this point as follows:

"[T]he trend in modern surveys is to make the data collected in the survey as widely available as possible, either ... by release of the raw data in a public use data file, or, more generally, by the construction of large scale survey data bases which integrate the data from many related surveys. These data bases are then used for a variety of secondary analyses by a wide range of analysts, most of whom will have had nothing to do with the original survey. ... For such a scenario, ... the internal consistency of survey estimates derived on a continuing basis from the survey data becomes paramount.

"Such consistency is easily achieved by using a method of estimation that allocates a unique weight, say \( w_i \), to each unit or case in the sample, with all survey estimates then being computed as weighted sums based on these case-weights. ... [A] natural question that arises is: How much efficiency (if any) is lost when a case-weighted strategy ... is used instead of the conventional ratio estimation strategy ... ?"

This was the question that Chambers set out to answer. Since each sample unit had to have a single sample weight it was necessary for all the relevant explanatory variables to figure in the equation for that weight simultaneously. He used two models for this purpose. The smaller one used five explanatory variables: namely the four factor-of-production variables mentioned above and an overall intercept. The larger model used eleven: the same four factor-of-
production variables plus seven zero-one indicators for the seven Australian Standard Industry Classification (ASIC) industries represented in the population. The larger model provided the stronger challenge and the comparisons presented here relate to that model only.

The five estimators that Chambers fitted his two models to were the following.

(1) The "standard" Generalized Regression Estimator, "GREG," (as defined by Särndal, Swensson and Wretman 1992, p225) namely

\[ \hat{T}_{GREG}(y) = 1_n' \Pi_i^{-1} y + (1_n' X - 1_n' \Pi_i^{-1} X_i) [X' Z_i \Pi_i^{-1} X_i]^{-1} X'_i Z_i \Pi_i^{-1} y_i. \]

[\( \hat{T}_{COSCAL}(y) \) has almost the same form as this. The only difference is that, where \( \hat{T}_{GREG}(y) \) uses \( Z_i \Pi_i^{-1} \), \( \hat{T}_{COSCAL}(y) \) uses \( Z_i (\Pi_i^{-1} - I_n) \).]

(2) The best linear unbiased predictor "BLUP," specified by

\[ \hat{T}_{BLUP}(y) = 1_n' y + (1_n' X - 1_n' X_i) [X' A_i^{-2} X_i]^{-1} X'_i A_i^{-2} y_i. \]

(3) A ridge estimator, "RIDGE," that eliminates unacceptable weights by moving along the ridge between the fully model-based "BLUP" and a sum of seven extension estimators, one for each ASIC industry.

(4 & 5) Two estimators ("NWD3" and "NWDAR3") that involve Nadaraya-Watson nonparametric adjustments to the weights for the estimator "RIDGE".
As a supplement to Chambers' study, the Cosmetic Calibration estimator, "COSCAL," was calculated for Sets 2-6. (For Set 1 "COSCAL" is identical with "GREG.") Many of the sample weights (the \( w_j \) of Section 6 above) were less than unity in the first instance, but wherever a sample contained one or more units with such unacceptably small weights, each of the corresponding \( \pi_j \) values was (by a convenient fiction) set equal to one and the calculation was then repeated. The presence of the matrix
\[
[X'(\Pi_s^{-1} - I_s)^{-1}Z_s'X_s]^{-1}
\]
in the formula for the sample weights vector
\[
w_j = \sigma_j + (\Pi_s^{-1} - I_s)^{-1}Z_s'X_s[X'(\Pi_s^{-1} - I_s)^{-1}Z_s'X_s]^{-1}[T(X) - X'\sigma]
\]
then ensured that wherever \( \pi_j \) had been set equal to unity, the corresponding \( w_j \) would also be unity. [Although "GREG" and "COSCAL" differ only in that "COSCAL" uses \( \Pi_s^{-1} - I_s \) where "GREG" uses \( \Pi_s^{-1} \), this difference is quite crucial when it comes to the elimination of unacceptable weights. For "COSCAL", the relevant component of \( \Pi_s^{-1} - I_s \) is equal to zero, so minimization of the distance function leaves any weight that has been set to unity unchanged. For "GREG," however, the corresponding component of \( \Pi_s^{-1} \) is equal to unity, so the weight can still be changed during recalculation.]

Between two and six iterations of this elimination procedure were typically required before all the 100 weights for a particular sample came out as greater than or equal to one, but occasionally one or more of the 100 weights could not be forced to take any value in this acceptable range. In these instances, only a few farms had been selected for a particular industry (usually the Dairy industry) and all those farms were larger than the average size for the industry
in question. It was then logically impossible to calculate any set of all positive weights defining an estimator that was calibrated both on the number of farms for that industry and on its total size measure. Samples selected with inclusion probabilities that varied rapidly with size were more subject to this problem than those for which the variation was slower.

The actual extent of the problem is indicated in Table A.1. Two samples in Set 2 each contained a single unacceptable weight. For Set 3, eight of the 500 samples contained one or more unacceptable weights; four contained one such weight and the other four contained two. Set 4 contained no unacceptable weights. A single sample in Set 5 contained two unacceptable weights. The procedure broke down completely only for Set 6. After two iterations of the elimination procedure, 45 out of the 500 samples still contained negative weights, a total of 154 between them, and seven of those 45 samples had already been identified as incorrigible. It seems probable that there was less of a problem for Set 3 than for Set 5 because Set 3 had no inclusion probabilities that were close to but not equal to one. No corresponding calculations have been made for the smaller model, but it is reasonable to suppose that the magnitude of the problem would be much less than for the larger one.
Table A.1. Progressive elimination of unacceptable COSCAL case-weights

<table>
<thead>
<tr>
<th>Sample Set</th>
<th>Iteration number</th>
<th>Number of samples with case-weights &lt;1</th>
<th>Incorrigible samples detected</th>
<th>Number of case-weights &lt;1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 2</td>
<td>0</td>
<td>277</td>
<td>0</td>
<td>496</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>85</td>
<td>0</td>
<td>127</td>
</tr>
<tr>
<td>Compromise allocation</td>
<td>2</td>
<td>18</td>
<td>0</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>16</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Set 3</td>
<td>0</td>
<td>226</td>
<td>0</td>
<td>701</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>100</td>
<td>1</td>
<td>303</td>
</tr>
<tr>
<td>Optimal allocation</td>
<td>2</td>
<td>48</td>
<td>1</td>
<td>134</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>27</td>
<td>4</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>13</td>
<td>7</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8</td>
<td>7</td>
<td>39</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>39</td>
</tr>
<tr>
<td>Set 4</td>
<td>0</td>
<td>188</td>
<td>0</td>
<td>322</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>55</td>
<td>0</td>
<td>80</td>
</tr>
<tr>
<td>Allocation</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>$\propto$ DSE$^{0.60}$</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Set 5</td>
<td>0</td>
<td>204</td>
<td>0</td>
<td>341</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>51</td>
<td>0</td>
<td>77</td>
</tr>
<tr>
<td>Allocation</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>$\propto$ DSE$^{0.75}$</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Set 6</td>
<td>0</td>
<td>187</td>
<td>0</td>
<td>592</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>96</td>
<td>1</td>
<td>229</td>
</tr>
<tr>
<td>Allocation</td>
<td>2</td>
<td>46</td>
<td>6</td>
<td>154</td>
</tr>
<tr>
<td>$\propto$ DSE$^{0.90}$</td>
<td>Further analysis abandoned</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.2 sets out a comparison of the initial incidences of unacceptably small weights for the estimators "GREG", "BLUP" and "COSCAL". The corresponding final incidence for "BLUP" is uniformly zero, but the estimator is then no longer "BLUP" but "RIDGE". No attempt was made to reduce the incidence of unacceptably small weights for "GREG", as the procedure was clearly going to be far less effective than it was for "COSCAL". For Set 2 the
The initial incidence of such small weights is substantially less for "BLUP" than it is for "GREG" or "COSCAL". For Set 3, however, the initial incidence for "COSCAL" is substantially smaller than that for "GREG" or "BLUP". The final incidences were calculated only for "COSCAL" and, as already noted, even then not for Set 6 (\( \pi_j \propto DSE_j^{0.90} \)), as the number of negative case-weights in incorrigible samples was already unacceptably large after two iterations.

Table A.2. Percentages of samples containing case-weights less than unity

<table>
<thead>
<tr>
<th>Sample Set</th>
<th>GREG</th>
<th>BLUP</th>
<th>COSCAL Initial</th>
<th>COSCAL Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1 (srswor)</td>
<td>77 (4.27)</td>
<td>77 (4.27)</td>
<td>77 (4.27)</td>
<td>n.c.</td>
</tr>
<tr>
<td>Set 2 (Compromise)</td>
<td>53 (1.83)</td>
<td>20 (1.83)</td>
<td>55 (1.79)</td>
<td>0.4 (1.50)</td>
</tr>
<tr>
<td>Set 3 (Optimal)</td>
<td>94 (9.73)</td>
<td>93 (5.87)</td>
<td>45 (3.10)</td>
<td>1.6 (4.88)</td>
</tr>
<tr>
<td>Set 4 (( \propto DSE^{0.60} ))</td>
<td>n.c.</td>
<td>n.c.</td>
<td>38 (1.71)</td>
<td>0.0</td>
</tr>
<tr>
<td>Set 5 (( \propto DSE^{0.75} ))</td>
<td>n.c.</td>
<td>n.c.</td>
<td>41 (1.67)</td>
<td>0.2 (2.00)</td>
</tr>
<tr>
<td>Set 6 (( \propto DSE^{0.90} ))</td>
<td>n.c.</td>
<td>n.c.</td>
<td>37 (3.17)</td>
<td>&gt;1.4</td>
</tr>
</tbody>
</table>

Numbers in parentheses are the average numbers of case-weights less than unity in samples containing at least one such case-weight.

Root mean squared errors (RMSEs) were also obtained for "COSCAL", both before and after the unacceptably small sample weights had been eliminated as far as possible.
Table A.3. RMSEs of the estimated population means of survey variables (expressed as percentages of the corresponding population values).

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Wheat income</th>
<th>Beef income</th>
<th>Sheep income</th>
<th>Dairy income</th>
<th>Total income</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATIO</td>
<td>14.7</td>
<td>28.9</td>
<td>19.1</td>
<td>14.4</td>
<td>16.7</td>
</tr>
<tr>
<td>GREG</td>
<td>13.6</td>
<td>26.1</td>
<td>17.0</td>
<td>15.0</td>
<td>17.3</td>
</tr>
<tr>
<td>BLUP</td>
<td>13.6</td>
<td>26.1</td>
<td>17.0</td>
<td>15.0</td>
<td>17.3</td>
</tr>
<tr>
<td>RIDGE</td>
<td>15.7</td>
<td>23.6</td>
<td>16.0</td>
<td>17.1</td>
<td>15.7</td>
</tr>
<tr>
<td>NWD3</td>
<td>15.0</td>
<td>22.1</td>
<td>15.9</td>
<td>17.5</td>
<td>14.6</td>
</tr>
<tr>
<td>NWD3AR</td>
<td>14.5</td>
<td>22.4</td>
<td>15.6</td>
<td>17.0</td>
<td>14.7</td>
</tr>
</tbody>
</table>

Set 1 (srswor)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Wheat income</th>
<th>Beef income</th>
<th>Sheep income</th>
<th>Dairy income</th>
<th>Total income</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATIO</td>
<td>10.0</td>
<td>11.6</td>
<td>15.5</td>
<td>19.2</td>
<td>8.3</td>
</tr>
<tr>
<td>GREG</td>
<td>9.9</td>
<td>11.9</td>
<td>14.8</td>
<td>20.3</td>
<td>8.4</td>
</tr>
<tr>
<td>BLUP</td>
<td>10.8</td>
<td>12.8</td>
<td>14.3</td>
<td>20.5</td>
<td>8.9</td>
</tr>
<tr>
<td>RIDGE</td>
<td>13.2</td>
<td>13.0</td>
<td>15.6</td>
<td>23.1</td>
<td>9.8</td>
</tr>
<tr>
<td>NWD3</td>
<td>10.5</td>
<td>11.5</td>
<td>14.1</td>
<td>19.8</td>
<td>8.1</td>
</tr>
<tr>
<td>NWD3AR</td>
<td>10.5</td>
<td>11.6</td>
<td>14.1</td>
<td>19.7</td>
<td>8.1</td>
</tr>
<tr>
<td>COSCAL:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial</td>
<td>9.9</td>
<td>12.1</td>
<td>14.8</td>
<td>20.3</td>
<td>8.4</td>
</tr>
<tr>
<td>Final</td>
<td>9.9</td>
<td>12.0</td>
<td>14.8</td>
<td>21.1</td>
<td>8.4</td>
</tr>
</tbody>
</table>

Set 2 (Compromise)

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Wheat income</th>
<th>Beef income</th>
<th>Sheep income</th>
<th>Dairy income</th>
<th>Total income</th>
</tr>
</thead>
<tbody>
<tr>
<td>RATIO</td>
<td>10.1</td>
<td>10.1</td>
<td>15.9</td>
<td>25.7</td>
<td>7.9</td>
</tr>
<tr>
<td>GREG</td>
<td>11.6</td>
<td>11.6</td>
<td>17.4</td>
<td>32.3</td>
<td>8.4</td>
</tr>
<tr>
<td>BLUP</td>
<td>11.9</td>
<td>11.1</td>
<td>16.4</td>
<td>32.1</td>
<td>8.0</td>
</tr>
<tr>
<td>RIDGE</td>
<td>23.5</td>
<td>9.6</td>
<td>21.3</td>
<td>47.8</td>
<td>11.9</td>
</tr>
<tr>
<td>NWD3</td>
<td>12.5</td>
<td>9.1</td>
<td>15.6</td>
<td>30.7</td>
<td>7.3</td>
</tr>
<tr>
<td>NWD3AR</td>
<td>12.9</td>
<td>8.9</td>
<td>15.7</td>
<td>31.5</td>
<td>7.3</td>
</tr>
<tr>
<td>COSCAL:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial</td>
<td>11.6</td>
<td>11.4</td>
<td>17.6</td>
<td>32.5</td>
<td>8.3</td>
</tr>
<tr>
<td>Final</td>
<td>14.6</td>
<td>11.6</td>
<td>18.1</td>
<td>41.4</td>
<td>8.7</td>
</tr>
</tbody>
</table>

n.c. not calculated. Interim. intermediate (after two iterations)
Table A.3 (concluded) RMSEs of the estimated population means of survey variables (expressed as percentages of the corresponding population values).

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Wheat income</th>
<th>Beef income</th>
<th>Sheep income</th>
<th>Dairy income</th>
<th>Total income</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set 4 ($= DSE^{0.60}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COSCAL: Initial</td>
<td>8.9</td>
<td>13.0</td>
<td>13.5</td>
<td>20.2</td>
<td>8.9</td>
</tr>
<tr>
<td>Final</td>
<td>8.8</td>
<td>11.4</td>
<td>13.6</td>
<td>20.1</td>
<td>7.9</td>
</tr>
<tr>
<td>Set 5 ($= DSE^{0.75}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GREG</td>
<td>9.7</td>
<td>10.3</td>
<td>14.4</td>
<td>20.4</td>
<td>7.5</td>
</tr>
<tr>
<td>COSCAL: Initial</td>
<td>9.7</td>
<td>10.4</td>
<td>14.5</td>
<td>20.4</td>
<td>7.5</td>
</tr>
<tr>
<td>Final</td>
<td>9.7</td>
<td>10.4</td>
<td>14.4</td>
<td>21.4</td>
<td>7.5</td>
</tr>
<tr>
<td>Set 6 ($= DSE^{0.90}$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COSCAL: Initial</td>
<td>11.0</td>
<td>11.4</td>
<td>16.9</td>
<td>29.8</td>
<td>8.5</td>
</tr>
<tr>
<td>Interm.</td>
<td>10.8</td>
<td>10.9</td>
<td>17.2</td>
<td>57.8</td>
<td>8.3</td>
</tr>
</tbody>
</table>

Table A.3 contains a comparison between these RMSEs and those reported in Chambers (1996) for the other estimators.

The optimum inclusion probabilities (as indicated by the position of the minimum RMSE) vary depending on what is being estimated. Set 4 with its $\pi_j \approx DSE_j^{0.60}$ seems close to being optimal for Wheat Income, Sheep Income and Dairy Income. Set 5 with $\pi_j \approx DSE_j^{0.75}$ is best for Beef Income and Total Income (which in this population is heavily dependent on Beef Income). Set 6 with $\pi_j \approx DSE_j^{0.90}$ is clearly too heavily slanted towards the selection of large farms, as is the "Optimal Allocation" of Set 1. The "Compromise Allocation" of Set 2 is comparable with Sets 4 and 5, but not quite as good as either. This is in keeping with the observation at the start of Section 1 [under
"Advantage (i)" that under \( \xi[0,1;v(x)] \) the optimum inclusion probabilities are proportional to \( \{v(x_j)\}^{0.5} \) across the entire population. For Wheat, Sheep and Dairy Income \( \{v(x_j)\}^{0.5} \) would appear to be roughly proportional to \( \text{DSE}^{0.60}_j \) and for Beef and Total Income more nearly proportional to \( \text{DSE}^{0.75}_j \).

Selections made with probabilities proportional to \( \text{DSE}^{0.60}_j \) and \( \text{DSE}^{0.75}_j \) are also very suitable for eliminating case-weights less than unity, but since this fact has nothing directly to do with the behaviour of any of the survey variables, it seems to be no more than a happy coincidence. \( \text{DSE}^{0.60}_j \) is the best over-all choice when both these considerations are taken into account, but if the smaller model had been used, with its lesser proclivity to produce case weights less than one, \( \text{DSE}^{0.75}_j \) might have been virtually as good.

As might be expected, the RMSEs obtained from the initial "COSCAL" are very similar to those obtained from "GREG". Little change is seen in the RMSE when the initial and final RMSEs for "COSCAL" are compared for the near-optimal sample Sets 2, 4 and 5. The somewhat surprising improvement for Beef and to a lesser extent Total Income is probably explicable in terms of some of the larger Beef Industry farms being atypical and therefore best prevented from contributing to \( \hat{\beta}_{\text{COSCAL}} \). The deterioration in the Dairy Income estimates is almost certainly due to the small numbers of Dairy Industry farms selected (particularly in Sets 3 and 6) and the consequently rapid rise in RMSE that occurred as farms with unacceptable case-weights were deleted and the effective sample size was decreased.
Turning to the choice of estimator, no one out of "RATIO", "GREG", "BLUP" and "COSCAL" has a consistent edge over any of the others on the sole ground of low RMSE. "RIDGE" is generally inferior, as might be expected on account of its model bias, and the two Nadaraya-Watson estimators are generally superior, as might also be expected on account of their nonparametric calibration. However, their superiority is neither compellingly large nor consistent over all variables. For Wheat Income they are less efficient than "BLUP", "GREG", and "COSCAL".

If the choice is restricted to the three simplest estimators capable of producing unique case-weights, namely "BLUP", "GREG" and "COSCAL", then all three are comparable in accuracy but "BLUP" and "GREG" are inferior to "COSCAL" in the elimination of unacceptably small sample weights. It is true that "COSCAL" was not uniformly successful in eliminating such weights, but the test it faced was exceptionally severe. Eleven explanatory variables (seven categorical, four continuous) were used for samples of size \( n = 100 \), the totals of these explanatory variables included several linked pairs (such as number of Dairy cattle and number of Dairy farms) and the inclusion probabilities were, for two of the six sample sets, close to being proportional to the size measure (DSE). Such a stringent combination of requirements should seldom be encountered in normal survey practice. However, in circumstances where the explanatory variables include such linked pairs (each consisting of a size measure and a corresponding unit count) it would seem prudent to avoid using inclusion probabilities that increase rapidly with size, even at the expense of a moderate departure from the otherwise optimal rule that the \( \pi_j \) should be proportional to the \( a_j \).
IV. Some Final Comments

Since the appearance of Royall and Herson (1973a,b), the old choice between randomization and purposive selection, long supposed settled in favour of the former, has been recast as a new choice between design-based and model-based inference. For the former, randomization is a logical necessity. For the latter, purposive selection (renamed "balanced sampling") is highly desirable. Other things being equal, model-oriented designs have much to recommend them, including the greater efficiency of the estimator (for a given sample size) their robustness against the selection of an atypical sample, and the fact that model-based estimation is standard in statistics generally in a way that design-based (or randomization-based) inference is not.

Despite this, model-based inference has been slow to take off within the sampling community generally. There has been an understandable reluctance to abandon an approach that has served so well for so long, the usual mistrust of any new procedure, and a concern that small deviations from the assumed model could easily invalidate the estimates (not the case if the samples are appropriately balanced, a point unaccountably passed over by Hansen, Madow and Tepping 1983).

There have also been some more soundly based misgivings. As already mentioned, randomized sampling has achieved a degree of public acceptance that might take some time to displace, and is also appreciably simpler to use than balanced sampling. The latter point is especially telling if balance is
required on more than a single explanatory variable and/or if several samples are being selected from a single sample frame and are being rotated around it. However, the persuasiveness of the Representative Principle described in Part I, Section 2 may eventually be judged more fundamental than either of these. It can even be taken as justified from a model-oriented point of view if the model is taken to be

$$\xi_p: y_j = x_j' \beta + f(\pi_j, c_j) + \varepsilon_j; E_\varepsilon \varepsilon_j = 0, E_\varepsilon \varepsilon_j^2 = \sigma_j^2, E_\varepsilon \varepsilon_j \varepsilon_k = 0, \forall k \neq j,$$

where $f(\pi_j, c_j)$ is any unknown but deterministic function of the $\pi_j$ and of any known population matrix of characteristics, $C$. If these considerations are put together with the results presented in Part II, showing that randomized sampling can accommodate model-based inference simultaneously with design-based, and do so nearly as efficiently as balanced sampling itself, the notion of using model-based and design-based inference simultaneously in the context of a randomized sample may well be considered more than just an acceptable compromise.

This combined inference can be used either in the usual establishment survey context of a stratified random sample (Part II) or in the more general situation where the sample units have unequal inclusion probabilities (Part III). In the former case, generalized least squares regression (with a diagonal variance matrix) can be used on a stratum by stratum basis to achieve equality between the population total of each explanatory variable and the corresponding survey estimate of that total. This is the most important property provided by balanced sample strategies. The difference still remains that for a stratified
balanced sample the case-weights are equal within a stratum, whereas for a stratified random sample they vary, and the latter is on that account the less efficient design.

The extent of that inefficiency depends both on the number of explanatory variables and on the intrinsic variability of each, but if regression is carried out with a diagonal variance matrix it is an easy matter to calculate the sample weights, $w_{hj}$ within the $h$th stratum, and an even easier matter to calculate the consequent increase in variance using the ratio

$$R_h = n_h \frac{\sum_{j \in s_h} w_{jh}^2}{\left( \sum_{j \in s_h} w_{jh} \right)^2}$$

as a measure of it. [Since the difference $w_{hj} - (N/h/n_h)$ is of order $(N/h/n_h)^{1.5}$, $R_h - 1$ is of order $O(n^{-0.5})$, so the increase in variance can always be ignored if the sample is large enough.]

If the sample is selected using unequal inclusion probabilities, the required sample weights can be calculated using the formula

$$w_j = \pi_j + (\Pi^{-1} - I_n)^{-1}Z_i'X_n[X'_i(\Pi^{-1} - I_n)^{-1}Z_i'X_i]^{-1}[T(X) - X_i'\pi_i]$$

derived in Part III, Section 4. The corresponding "Cosmetic Calibrated" estimator of $\beta$, $\hat{\beta}_{coscal}$, can be expressed in either of the forms (vi) or (vii) of Part III, Section 5, and the estimator of $T(y)$, $\hat{T}_{coscal}(y)$, in any one of the
forms (i) to (v), which include two generalized regression estimators and two predictor forms.

For evaluation of the estimates from a particular sample, the model-variance of Part III, Section 6 is the appropriate one to estimate, but for an evaluation of a proposed sample design, the anticipated variance considered in Section 2 is preferable. This has the advantage over the design-variance that neither the anticipated variance itself nor its natural estimator depend on the second-order inclusion probabilities, \( \pi_{jk} \).

Another useful property of Cosmetic Calibration is that, except in extreme circumstances, case-weights less than one can easily be replaced by weights that are exactly one, and that in subsequent recalculations such weights automatically remain unchanged. Not only that, but these atypical units no longer impinge on the estimation of the regression coefficient. In ordinary circumstances, the problem of unacceptably small sample weights then virtually disappears, and this without any need to compromise on the requirement that the sample estimate of an explanatory variable should be equal to its population value. The empirical comparison (in Part III, Section 8) with the results obtained by Chambers (1996) indicate that the cosmetic calibration estimator is not only comparable in accuracy with the standard generalized regression estimator as defined by Särndal, Swensson and Wretman (1992) but also comparable with the best linear unbiased predictor itself, better than the standard ridge estimator, and only marginally worse than the two Nadaraya-Watson nonparametric modifications of that estimator.
There is nothing to prevent Nadaraya-Watson modifications being made to the cosmetic calibration weights. Accuracy would presumably be improved for Beef Income and Total Income by following the non-linearities, but the calibration property would be sacrificed and the estimates for Wheat Income could be expected to suffer. However, a more promising opportunity for improvement might be to substitute linear dependence of Beef Income on Numbers of Beef Cattle by a nonlinear one. Chambers, Dorfman and Wehrly (1993) found a surprisingly good linear relationship between log(Beef Income + $1) and log[log(Numbers of Beef Cattle)]. This implies proportionality between (Beef Income + $1) and log(Numbers of Beef Cattle), or for practical purposes simply between Beef Income and log(Numbers of Beef Cattle).

Unfortunately, the relationship was found only for farms with 50 or more beef cattle. No information is provided on a possible relationship for farms with fewer than 50 beef cattle, but with if the same proportionality happens to hold for them, the adjustment for nonlinearity is simple. For the particular population of 904 under study here, all that is then needed is to replace the explanatory variable Numbers of Beef Cattle by log(Numbers of Beef Cattle). It is, of course, likely that this particular proportionality will be less close for small beef herds, but the possibility that some (probably more complex) overall proportionality might be discovered seems worth exploring.

For other populations, the population total for the replacement explanator might not be known. If individual Numbers of Beef Cattle, and therefore also individual values of the replacement explanator, were known only for the units
in sample, it might be helpful to estimate the required population total using a cosmetic estimator calibrated both on Numbers of Beef Cattle and on an indicator variable for Farm With Beef Cattle. This also requires further investigation.

The motivation underlying the original research presented in Parts II and III was in the first instance only to define a class of estimators that could be regarded as valid regardless of whether the approach used was the design-oriented one pioneered by Neyman or the model-oriented one championed by Royall and his co-authors. It was seen from the beginning that this would result in a liberation of unequal probability sampling from the tyranny of the $\pi_{\mu}$—even the concept of "anticipated variance" itself (that provided the escape from the $\pi_{\mu}$) only has meaning if both approaches are used simultaneously—but some of the other results presented in those two Parts were less predictable.

These include the near-equivalence of the stratified balanced and stratified random sampling designs discussed in Part II, the complete equivalence of the Cosmetic and Calibration approaches in Part III (in retrospect explained easily enough by the fact that Cosmetic Calibration is the intersection of them), the consequent array of equivalent estimators presented in Section 5 of Part III, and, least expected of all, the remarkably easy way in which the problem of unacceptably small case-weights could be circumvented.

Another point, not mentioned earlier, relates to something that has been an increasingly common practice in recent years, namely the simultaneous use of
design-based estimation for large domains and synthetic estimation for small domains in the same study. Synthetic estimation being essentially model-based, it has been necessary to use a "forcing step" to adjust the totals of small domain estimates to equal the design-based estimate of each corresponding large domain. Cosmetic calibration being model-based and design-based simultaneously, the small domain synthetic estimates sum automatically to the large domain total. The elimination of the "forcing step" does not save a great deal of work, but it does mean a small step forwards in elegance.

In summary, the combination of design-based and model-based inference not only permits the choice of the more appropriate of two statistics or the more convenient of two procedures; it also provides insights and options that neither could produce by itself. In its own small way, therefore, it would seem that the bringing together of the two approaches is leading to the same kinds of "emergent phenomena" as the new science of Complexity is now finding in physics, biology, economics, computer science and many other fields (Lewin 1993, Wardrop 1993).

A final caution: Topic A has dealt only with business surveys. Surveys of individuals and households commonly involve two or more stages of sampling, and this introduces a complication into the reconciliation of design-based with model-based inference that has not so far been seriously addressed. I have in mind to attempt this next year, even if only to ensure that small area synthetic estimates can add automatically to the design-based ones customarily used for large areas.
REFERENCES


Godambe, V.P. (1955), "A Unified Theory of Sampling from Finite


Isaki, C.T., and Fuller, W.A. (1982), "Survey Design under the Regression


TOPIC B

Frequentist and Bayesian Tests of a Precise Null Hypothesis against a Composite Alternative

"The issue is simply that $p$ gives a very misleading impression to the validity of $H_0$, from almost any evidentiary viewpoint."

Berger and Sellke (1987, p112)

"... a Bayes factor approaching infinity does not mean strong support for the null model $M_1$, but strong rejection of the diffuse prior model $M_2$."

Aitkin (1991, p115)

"It seems reasonable ... to conclude that that we must embrace subjective Bayesian analysis, in some form, to reach sensible conclusions about testing a point null."

Berger and Sellke (1987, p120)
I. An Empirical Bayesian Treasure Hunt in Flatland

Stone's (1976) Treasure Hunt in Flatland is re-examined. The uniqueness of the Classical solution is questioned, and it is shown in any case that an empirical Bayesian solution can be substantially superior to any Classical alternative. Further, the diagram supplied by Stone can be interpreted in two ways. The first interpretation leads without any serious complication to the conclusion that another set of three treasure locations (different from either of those considered by Stone) is almost certain to contain the treasure. The second interpretation, however, leads to serious complications (involving uniform and diffuse priors) and is dealt with separately in Part II.

The ambiguity in Stone's diagram concerns the direction in which the treasure layers traversed the loop in the north-west corner. If they traversed it clockwise, then they never took three consecutive "erratic steps" either to the right, or forwards or to the left. To lay the treasure at a particular one of the potential locations they would have to have taken three consecutive "erratic" steps to the left, an event with very small likelihood. The three remaining locations provide the optimal set required in Stone's formulation of the problem.

KEY WORDS: Coverage Probability; Empirical Bayes; Noninformative prior.
1. INTRODUCTION

Stone (1976) introduced a deceptively simple-looking problem designed to reveal the existence of a strong inconsistency between coverage probabilities on the one hand and posterior probabilities calculated using a uniform prior distribution on the other. The essential features of the problem are described in the following quotation:

"Since Abbott [1884] described Flatland there have been many changes. The whole of Flatland has been developed along the lines of a North American city; blocks, streets (north-south and east-west) and intersections ...

"A soldier and a woman ... wander erratically through the streets of Flatland, never immediately retracing their steps. As they proceed, they let out a taut continuous thread attached to [their initial starting point at intersection] O, which passes through each intersection visited by them between O and T, where T denotes the intersection where, after a stochastically long time, they come to a halt. At the centre of T, the soldier fixes the woman so that she can spin freely about her midpoint. [In Flatland, women are like short needles whose 'eyes' have migrated to their midpoints. (Stone's footnote)] He gives her a vigorous spin and observes which adjacent intersection, $T_a$ say, she points to when she has come to rest. Leaving an invisible but valuable treasure at $T$, the soldier and the woman proceed to $T_a$, possibly retracing the last step of their path between $O$ and $T$. In the case of retracing they take up
Strong Inconsistency from Uniform Priors

An Outcome of the Treasure Laying

Figure 1. A possible outcome of the treasure laying as supplied by Stone (1976)
the slack in the thread between $O$ and $T_a$ and, in any case, leave it attached to $T_a$.

"The figure [reproduced here as Figure 1] shows a possible outcome of the procedure in which the treasure was left at $T = T_a^E$, say, the intersection adjacent to and east of $T_a$. The final taut state of the thread between $O$ and $T_a$ removes any direct evidence of the location of $T$, and the inference problem ... is: 'Which of the possibilities \{$T_a^N, T_a^S, T_a^E, T_a^W$\} is $T$?''

The full formulation of Stone's problem involves also a Bayesian Statistician, $B$, and a Classical Statistician, $C$, who are allowed to traverse the path from $O$ to $T_a$ and are then independently given the task of choosing any set of three from the four possible sites for $T$. They are regarded as "winning"—again independently—if their chosen set includes $T$. (Their real task is therefore to identify the least probable location and exclude it from the chosen set.) The treasure laying and location choosing occur "repeatedly", so it is in principle sound to choose the locations for which the coverage probabilities are highest.

Stone's Bayesian Statistician chooses a uniform prior over all intersections in Flatland, notes that the likelihoods of travelling in the four possible directions from $T$ to $T_a$ are each equal to 1/4, and arbitrarily - "to keep things tidy" - excludes the one that did not extend the path. In Stone's diagram this was $T_a^E$. His Classical Statistician, more astutely, notes that if the path was extended the treasure would be located at $T_a^E$, and that the coverage probability of extending the path was 3/4. Consequently he chooses $T_a^E$ and,
again to keep things tidy, the two "around the corner" from $T^f_a$, thereby excluding $T^m_a$. Stone noted that for any finite path starting from $O$ the coverage probability that his Classical Statistician would win was $\geq 3/4$, whereas the similarly defined probability that his Bayesian Statistician would win was $\leq 1/4$. (He could in fact have said exactly $1/4$.)

The strong inconsistency that Stone saw as arising from the use of a uniform prior was, however, that his Bayesian Statistician's posterior probability of winning was $3/4$, whereas his coverage probability of winning was only $1/4$. He challenged his Bayesian contemporaries to produce a solution using a uniform prior that did not give rise to this strong inconsistency. Despite the fact that a Bayesian solution was offered that reproduced the one offered by Stone's Classical Statistician, it was generally accepted that his problem had raised difficulties for Bayesian statisticians and in particular that he had made his point regarding the dangers inherent in the use of improper priors.

Here I challenge that view. In Section 2 the validity of the Classical Solution is called into question. The solution to Stone's problem developed in the remainder of Part I is offered primarily as a demonstration that an empirical Bayesian analysis can improve on the Classical Solution quite markedly. In Part II a second possible interpretation of Stone's problem is used to show that improper priors can provide sensible posterior odds in a context where this has hitherto been regarded as impossible; namely, the test of a precise null hypothesis against a composite alternative. Part III elaborates some of the implications of this discovery.
2. THE CLASSICAL SOLUTION REVISITED

In order to understand Stone's Classical solution properly we will need to investigate what he meant by "wander erratically ... never immediately retracing their steps" and "'coverage' probabilities." Neither of these is defined in Stone's article and neither is entirely unambiguous, but the meaning of both can still be established within fairly narrow bounds by examining their context.

There are two possible meanings of "erratic wandering," one narrow and the other more general. The narrow interpretation is the only one that appears in the discussion (Dickey 1976, Box and Tiao 1976) but there is internal evidence that Stone intended the more general one to be understood. The narrow one is that the treasure layers maximized the entropy of their path by choosing equiprobably among a right turn, a left turn and a forward step at each intersection. We will refer to this interpretation as "conditional(ly) random" and denote it by CR.

The possible wider interpretation is that at each intersection the treasure layers made completely arbitrary decisions as to which of the three permissible directions they would take. CR is a special case of this, but if their path was not in fact generated by CR, its entropy is less than maximal, so the potential exists to detect structure in it and to make predictions more precise than would be possible under CR. Since CR is a special case of it, we
will adopt the more general interpretation on a tentative basis from this point onwards. CR will nevertheless continue to receive due attention.

Before mentioning "coverage' probabilities", Stone had already used \( \theta \) to "denote the thread between \( O \) and \( T \)," \( x \) to denote "that between \( O \) and \( T_a \) and \( x^N, x^S, x^E, x^W \) to "denote hypothetical taut threads (paths) that are identical with \( x \) except that they terminate at \( T_a^N, T_a^S, T_a^E \) and \( T_a^W \) respectively." The only four possible values of \( \theta \) were, therefore, \( x_N, x_S, x_E \) and \( x_W \). He first used the expression "coverage' probabilities" to introduce two of his equations as follows:

The following "coverage" probabilities are easily evaluated:

\[
P(C \text{ wins } \theta) \geq \frac{3}{4}, \quad (2.2)
\]
\[
P(B \text{ wins } \theta) \leq \frac{1}{4}, \quad (2.3)
\]

for all \( \theta \).

It is indeed easy enough to make out the gist of what Stone meant by these two equations. We may with some assurance interpret this passage as follows:

Whatever the true location of the treasure, a strategy that in the circumstances illustrated by Figure 1 chooses the set \( \{x^E, x^N, x^S\} \) will win at least three quarters of the time but one that chooses \( \{x^N, x^S, x^W\} \) will win at most one quarter of the time.
Figure 2. The four possible paths from $T_a^B$ to $T_a$.

Erratic steps shown as unbroken lines.
Random steps shown as dashed lines.
Stone could in fact have written (2.3) as \( P(B \text{ wins} | \theta) = \frac{1}{4} \), changing the interpretation of the last phrase to "will win one quarter of the time." This is probably no more than a misprint, but the way in which he uses \( \theta \) in (2.2) and (2.3) is a serious puzzle. If \( \theta \) takes the values \( x^N, x^E, x^W \) and \( x^W \) in turn, then the probabilities that \( C \) wins are 1, 1, 1, and 0, while the probabilities that \( B \) wins are 1, 1, 0 and 1 respectively. There must be some better way to convey the part of the message that we have interpreted as meaning "Whatever the true location of the treasure," than to say that (2.2) and (2.3) hold "for all \( \theta \)." However, until that way is found, it is not possible to specify precisely what is meant by "'coverage' probabilities."

We will endeavour to find such a way using Figure 2. This illustrates the paths that would need to be followed if the treasure were to have been buried at \( T^B_a, T^R_a, T^F_a \) and \( T^L_a \) respectively, where \( T^B_a \), denotes the intersection adjacent to \( T_a \) backwards along the thread, \( T^R_a \), the one arrived at by turning right, \( T^F_a \) the one forwards and \( T^L_a \) the one on the left. (In Figure 1 these are \( T^R_a, T^N_a, T^W_a \) and \( T^F_a \) respectively.) If the final random step was taken backwards (an event with objective probability 1/4, resulting in a shortening of the thread and hence three steps between \( T \) and \( T_a \)) that step random step must have started from one of \( T^R_a, T^F_a \) or \( T^L_a \), so the sum of the probabilities associated with the treasure being located at these three locations is at least, and in fact exactly equal to, an objective 1/4. However, the individual probabilities associated with the random step starting from \( T^R_a, T^F_a \) and \( T^L_a \) cannot be known, because up to though not including the last step, the probabilities were being determined by the unknown rules of "erratic
wandering." The remaining three individual objective probabilities, each of 1/4, are associated with the single step paths illustrating the possible right turn, forward step and left turn, each of which would have originated from $T_a^B$ and have lengthened the thread. The Classical Statistician's conclusion that $T_a^B$ is three times as good a choice as the other three combined is therefore dependent entirely on the known and objective probabilities derived from the randomness of the final step from $T$ to $T_a$. It is this exclusive dependence on these known objective probabilities that characterises the Classical solution and appears to supply the expression "'coverage' probabilities" with its intended meaning.

A possible objection to this solution is that the two statisticians were allowed to traverse the entire path of the thread, and hence the Classical Statistician should have known from which direction $T_a^B$ was approached. (In Stone's figure it is such that the next step—the one to $T_a$—requires a left turn.) That direction being known, the other two possible directions could be eliminated, reducing the coverage probability for $T_a^B$ to 1/4, or to 1/2 after normalization, the other 1/2 being associated with the union of $T_a^B$, $T_a^f$ and $T_a^l$.

If this objection is allowed, a "'coverage' probability" can only be specified in the context of what is known about the path. If all that is known is the location of $T_a$ and the direction that the thread takes backwards from it, the coverage probability associated with $T_a^B$ is 3/4, but if the location of the intersection visited immediately before $T_a^B$ is also known, the coverage probability is only 1/2.
If at this point any readers feel that something must be wrong and that the "real" probability of the treasure being located at $T_u^B$ must be greater than 1/2, they can be assured that their misgivings are almost certainly justified. However, the problem is not that the additional information about the path is irrelevant and should be ignored. The reverse is true. There is a great deal more information that the path can supply, and all of it is relevant. The real problem is that in concentrating attention exclusively on the objective probabilities supplied by the final random step, most of the relevant information is being discarded. The most straightforward means by which that information may be put to work is by an empirical Bayesian analysis, and this approach will be followed in Section 4 onward. As a preliminary to this, we will (in Section 3) examine the comments made by the Bayesian statisticians during the Discussion on Stone's paper. Some of the implications reinforce conclusions already reached in this Section.

3. REPRODUCING THE CLASSICAL SOLUTIONS USING BAYESIAN ARGUMENTS

Several Bayesian discussants to Stone's paper were able to produce a Bayesian solution equivalent to Stone's Classical solution (SC). The clearest description is that given by Dickey (1976). All, however, used the assumption that at every erratic step the probabilities of turning right, moving forward, and turning left were each equal to 1/3 (the "conditionally random" or CR assumption). Refering again to Figure 2, the likelihoods of the treasure being located at the possible sites were then
\[ \text{Lik} (T_a^B) = 1/4 \quad \text{and} \quad \text{Lik} (T_a^R) = \text{Lik} (T_a^F) = \text{Lik} (T_a^L) = 1/3 \times 1/3 \times 1/4 = 1/36. \]

Combining these likelihoods with a uniform prior over all intersections in Flatland, the posterior probabilities were

\[ \text{PostPr} (T_a^B) = 3/4 \quad \text{and} \quad \text{PostPr} (T_a^R) = \text{PostPr} (T_a^F) = \text{PostPr} (T_a^L) = 1/12. \]

SC is therefore the Bayes solution given CR, but the alternative Classical solution (AC) also has a Bayesian equivalent in the extreme but possible case where it can be predicted with probability one (given the rules of erratic wandering) that the next step—if any—after \( T_a^B \) will be to the observed \( T_a \), for then

\[ \text{Lik} (T_a^B) = 1/4 \quad \text{and} \quad \text{Lik} (T_a^R) + \text{Lik} (T_a^F) + \text{Lik} (T_a^L) = 1/12 \times 1/12 = 1/4 \]

and

\[ \text{PostPr} (T_a^B) = 1/2 \quad \text{with} \quad \text{PostPr} (T_a^R) + \text{PostPr} (T_a^F) + \text{PostPr} (T_a^L) = 1/2. \]

Hence no solution based entirely on the known objective probabilities from the random step has universal validity. The mechanism generating the erratic path cannot be ignored. (Paradoxically, in the Bayesian arguments the known objective probabilities cancel out.)
4. AN EMPIRICAL BAYESIAN APPROACH

The Bayesian arguments of the last Section depend critically on the probabilities of travelling right, forwards and left for the last two "erratic wandering" steps; the one from $T_*^B$ to $T_a$, and the other from $T_a$ to one of the three other possible treasure locations.

The CR assumption used in 1976 was unnecessary. The Flatland statisticians had been allowed to follow the entire observable path of the thread, and they could have used this to obtain more realistic estimates of the relevant likelihoods. The Classical Statistician would have had little incentive to do so, because the Classical solution was deliberately based only on known and objective probabilities, but a Bayesian statistician should naturally have been led to examine the path for such evidence.

Consider the particular path provided by Stone in Figure 1, and how it could have been used by a more sensible Bayesian statistician than Stone's. We will call this person "Barbara", after the mediæval mnemonic used to denote the simplest valid syllogism. (Abbott's narrator would not have approved of a woman becoming any sort of statistician, but of course he was a Square.)

For convenience we start Barbara off not at $O$ but at $T_a$, and allow her to trace the thread backwards. She takes $T_a$ as her point of reference, identifies "backwards" with "eastwards", and proceeds as follows. First for $T_a^E$, 

\( \text{Lik}(T^S_a) = Pr(t|x').1/4 \),

where \( Pr(t|x') \) denotes the probability that the treasure is laid at the end of the path \( x' \), and \( x' \) denotes that part of the thread which unambiguously delineates erratic movements (i.e., the entire path up to but not including the last observable step).

Next for \( T^N_a \),

\[ \text{Lik}(T^N_a) = \{ 1-Pr(t|x') \} Pr(LWT_a|x') \{ 1-Pr(t|x) \} Pr(RNT^N_a|x) Pr(t|x, RNT^N_a)/4, \]

where \( Pr(LWT_a|x') \) is the probability of a left turn west into \( T_a \), given \( x' \), \( x \) denotes the entire thread up to and including \( T_a \), \( Pr(RNT^N_a|x) \) is the probability of a right turn north into \( T^N_a \) given \( x \) and \( Pr(t|x, RNT^N_a) \) is the probability of leaving the treasure at the end of the path consisting of \( x \) plus the right turn north into \( T^N_a \). The expressions for the remaining two likelihoods are analogous:

\[ \text{Lik}(T^W_a) = \{ 1-Pr(t|x') \} Pr(LWT_a|x') \{ 1-Pr(t|x) \} Pr(FWT^W_a|x) Pr(t|x, FWT^W_a)/4, \]

and

\[ \text{Lik}(T^S_a) = \{ 1-Pr(t|x') \} Pr(LWT_a|x') \{ 1-Pr(t|x) \} Pr(LST^S_a|x) Pr(t|x, LST^S_a)/4. \]
Barbara knows about the "stochastically long time" it took to reach $T$, so assumes the five treasure laying probabilities are all small compared with unity. Thinking along the same lines as Lindley (1976) she concludes that although they must eventually decrease as the number of steps increases, they may also be considered as nearly equal. Finally, the factors $1/4$ can (as in Section 3) be cancelled out. Her simplified likelihoods are

$$Lik^*(T_0^E) = 1,$$

$$Lik^*(T_0^N) = Pr(LWT_0|\lambda^*).Pr(RNT_0^N|\lambda),$$

$$Lik^*(T_0^W) = Pr(LWT_0|\lambda^*).Pr(FWT_0^W|\lambda)$$

and

$$Lik^*(T_0^S) = Pr(LWT_0|\lambda^*).Pr(LST_0^S|\lambda).$$

Four unknown probabilities remain to be determined, but she can gather information regarding them from the remainder of the path.

5. THE CLOCKWISE SCENARIO

We who have to rely on Stone's diagram (Figure 1) cannot tell whether the treasure layers travelled clockwise or anticlockwise around the loop in the north-west corner, but for Barbara the thread is as she finds it. Since,
however, her problem is much simpler if the treasure layers followed the loop in a clockwise fashion, we shall consider that scenario first.

Of the 120 steps, the first 119 were unambiguously travelled in an erratic fashion. The 120th might have been erratic or random. She notes that the first 119 were travelled East, North, West or South in the following (row-wise) order:

\[
\begin{align*}
NWNEN & \quad ENNWN & \quad WNEEN & \quad WNNNE & \quad SEENN \\
WNNES & \quad SEESE & \quad SWSSS & \quad EENES & \quad SWSEE \\
ENNNE & \quad SENNN & \quad WNNWN & \quad EENEE & \quad SESSE \\
SENE & \quad NNEEN & \quad ESSSW & \quad SESWS & \quad WNWSS \\
SWSEE & \quad NEEES & \quad SESEE & \quad SEEN.
\end{align*}
\]

The frequencies of occurrence are \( E \ 41, \ N \ 33, \ W \ 14 \) and \( S \ 31 \). The CR hypothesis is already questionable. There may be a propensity to drift eastwards. Barbara constructs the Markov Chain transition probabilities of Table B.1 to reflect both this drift and the fact that the pair never retraced their steps.
Table B.1. Basic Transition Probabilities for the Treasure Layers

<table>
<thead>
<tr>
<th>Previous Step</th>
<th>$E$</th>
<th>$N$</th>
<th>$W$</th>
<th>$S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
<td>4/10</td>
<td>3/10</td>
<td>0/10</td>
<td>3/10</td>
</tr>
<tr>
<td>$N$</td>
<td>4/9</td>
<td>3/9</td>
<td>2/9</td>
<td>0/9</td>
</tr>
<tr>
<td>$W$</td>
<td>0/8</td>
<td>3/8</td>
<td>2/8</td>
<td>3/8</td>
</tr>
<tr>
<td>$S$</td>
<td>4/9</td>
<td>0/9</td>
<td>2/9</td>
<td>3/9</td>
</tr>
</tbody>
</table>

Under Table B.1, the relative frequencies of travel $E$, $N$, $W$ and $S$ are 40:27:16:27 (summing to 110). 119 unambiguously erratic steps were observed, but the first step had no "previous step", so the 118 "following" erratic steps are distributed $E$ 41, $N$ 32, $W$ 14 and $S$ 31. For Table B.1 the expected numbers of steps are (to one decimal) 42.9, 29.0, 17.2 and 29.0.

Barbara knows that an exclusive reliance on likelihood ratio tests can lead to the choice of a badly fitting model over an even worse fitting model. She therefore does a standard goodness-of-fit test before she proceeds any further. The value 1.13 for chi-square with three d.f. reassures her to the extent that there is nothing in the marginal frequencies that she has summarized to suggest that Table B.1 is an obviously inadequate model. The corresponding value for the CR hypothesis (that the probabilities of travel in the three permitted directions are always 1/3 each) is 12.92, which
Frequentists would describe as "significant at the one per cent level", but Barbara distrusts the Frequentist practice of counting the improbabilities of events that have not occurred. This makes them overestimate the significance of their tests (Berger and Delampady, 1987, Subsection 4.5). Reassured by the low chi-square value for Table B.1, she conducts a likelihood ratio test of its transition probabilities against the CR hypothesis. The resulting Bayes Factor is

$$\frac{(4/10)^{14}(3/10)^{27}(4/9)^{22}(3/9)^{22}(2/9)^{14}(3/8)^{14}}{(1/3)^{118}} \equiv 31.4.$$ 

This is not overwhelming evidence against CR, but since any weight given to that model would merely tend to make the three least probable locations more equiprobable, Barbara decides to ignore it for practical purposes and consider only other possible deviations from the model of Table B.1. (We may note in passing that if Table B.1 is correct, the decision set chosen by Stone's Classical Statistician is optimal. However, if Barbara happens to choose that set, she will do so on better grounds than merely because $T_w^W$ and $T_w^S$ are "around the corner" from $T_w^F$, and $T_w^W$ is not.)

To detect possible deviations from Table B.1, Barbara next records whether, at each step following the first, the treasure layers travelled Right, Forward or Left. She finds the following:
The marginals are $R\ 41,\ F\ 36,\ L\ 41$. Starting from scratch, Barbara would have carried out a goodness of fit test on the CR hypothesis and have obtained the result that chi-square with two d.f. was 0.42. But with Table B.1 as her basic model, the relative frequencies 18:19:18 become her point of reference. Her expected frequencies are then $R\ 38.6,\ F\ 40.8$ and $L\ 38.6$. The resulting chi-square is 0.85, somewhat bigger but still not significantly large.

However, the sequences $RRR,\ FFF$ and $LLL$ are remarkable by their absence. Their expected numbers, given Table B.1, would have been $RRR\ 3.9,\ FFF\ 5.2$ and $LLL\ 3.9$, totalling 12.9. Their observed numbers are all zero. Barbara suspects that the erratic mechanism is inhibiting the third $R,\ F$ or $L$ following any sequence of two, decreasing the incidence of all runs of two by one third. Table B.1 then leads her to expect $RR\ 12.0,\ FF\ 14.4$ and $LL\ 12.0$, totalling 38.4. The observed numbers are $RR\ 11,\ FF\ 8$ and $LL\ 7$, totalling 26. This reduction is so close to the theoretical value of one third that Barbara goes straight to the likelihood ratio test. Her null hypothesis is Table B.1. Her alternative is Table B.1 supplemented with Table B.2 following $RR$, Table B.3 following $FF$ and Table B.4 following $LL$. These last three tables are constructed by taking Table B.1, setting the probability of a third move similar
to the last two at zero, and prorating the remaining two probabilities so as to sum to unity.

Table B.2. Transition Probabilities Following Two Successive Right Turns

<table>
<thead>
<tr>
<th>Following Step</th>
<th>E</th>
<th>N</th>
<th>W</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E</strong></td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Previous</td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Step</td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>W</strong></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>S</strong></td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table B.3. Transition Probabilities Following Two Successive Forward Steps

<table>
<thead>
<tr>
<th>Following Step</th>
<th>E</th>
<th>N</th>
<th>W</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E</strong></td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Previous</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td><strong>N</strong></td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Step</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>W</strong></td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>S</strong></td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table B.4. Transition Probabilities Following Two Successive Left Turns

<table>
<thead>
<tr>
<th></th>
<th>Following Step</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E$</td>
<td>$N$</td>
<td>$W$</td>
<td>$S$</td>
</tr>
<tr>
<td>$E$</td>
<td>$\frac{4}{7}$</td>
<td>0</td>
<td>0</td>
<td>$\frac{3}{7}$</td>
</tr>
<tr>
<td>$N$</td>
<td>$\frac{4}{7}$</td>
<td>$\frac{3}{7}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$W$</td>
<td>0</td>
<td>$\frac{3}{5}$</td>
<td>$\frac{2}{5}$</td>
<td>0</td>
</tr>
<tr>
<td>$S$</td>
<td>0</td>
<td>0</td>
<td>$\frac{2}{5}$</td>
<td>$\frac{3}{5}$</td>
</tr>
</tbody>
</table>

The Bayes Factor in favour of her alternative hypothesis is a function only of the likelihood ratios for the 26 steps following runs of two:

$$(10/7)^8(9/6)^6(8/5)^3 \cong 31,100.$$ 

This is highly significant by any reasonable criterion. Barbara, however, wishes to be on the safe side. Her subjective prior odds favouring Table B.1 alone over her more elaborate alternative hypothesis were not negligible. She reckons that they would have been more than 9:1 on but less than 99:1 on. She decides she will, for practical purposes, use posterior odds of 999:1 on the more elaborate hypothesis.

She is now ready to work out her optimal decision set. Had Table B.1 been acceptable by itself she would have used
\[ Pr(LWT^a|x^e) = 2/9, \]
\[ Pr(RNT^N_a|x) = 3/8, \]
\[ Pr(FWT^w_a|x) = 2/8, \]

and
\[ Pr(LST^S_a|x) = 3/8. \]

These values would have yielded
\[ PostPr(T^E_a) = 36/44 = 9/11, \]
\[ PostPr(T^N_a) = 3/44, \]
\[ PostPr(T^w_a) = 2/44, \]

and
\[ PostPr(T^S_a) = 3/44. \]

The posterior probability that the treasure was laid at \( T^E_a \) is somewhat larger than the Classical value of \( 3/4 \). This is because of the eastward drift implied by Table B.1. The relative improbability of \( T^N_a \) has the same cause.

Barbara, however, noting that the last two erratic steps before \( T^E_a \), \( T^N_a \) and \( T^S_a \) would in each case have been LL, now sees it as far more probable that the last three likelihoods should be taken from Table B.4 rather than Table B.1. This would imply
\[ Pr(LWT^a|x^e) = 2/9, \]
\[ Pr(RNT^N_a|x) = 3/5, \]
\[ Pr(FWT^w_a|x) = 2/5, \]
and

$$Pr(LS T^S_a | x) = 0.$$ 

These values would have yielded

$$PostPr(T^S_a) = \frac{45}{55} = \frac{9}{11},$$

$$PostPr(T^W_a) = \frac{6}{55},$$

$$PostPr(T^W_b) = \frac{4}{55},$$

and

$$PostPr(T^S_b) = 0.$$ 

The posterior probability associated with $T^S_a$ remains at $\frac{9}{11}$, but the other three are substantially changed and the least probable location for the treasure is now $T^S_a$. Barbara gives (partly subjective) weights $0.001$ to the first set of probabilities and $0.999$ to the second, but $T^S_a$ is still obviously the one to be excluded from her set of three. Her subjective probability of winning, if she excludes $T^S_a$, is between $0.9998$ and $0.99998$.

Cautiously, however, she carries out a sensitivity analysis before making her final decision. She notes that if she were to give weight $\frac{16}{21}$ to Table B.1 alone and weight $\frac{5}{21}$ to the alternative, she would be indifferent between excluding $T^S_a$ and $T^W_a$. Could she conceivably give that much weight to Table B.1? She would not be prepared to consider such an extreme prior odds ratio as $100,000:1$ on Table B.1 against her more elaborate hypothesis, but she does wonder what would have happened had she been prepared to consider a
mixed model with the unknown weight to be applied to Tables B.2, B.3 and B.4 as a parameter to be estimated.

She feels uncomfortable about this. It is not the possibility that the mixed model might be relatively probable that worries her. No runs of three were found in a situation where with Table B.1 alone one would expect 12 or 13, so it seems implausible that the parameter estimate for the alternative hypothesis could ever be in the region of 16/21. It is rather that the choice of reference prior for the test of a precise null against a composite alternative is such a hotly debated topic (Kass and Wasserman, 1995; O'Hagan, 1995; Berger, 1995; Aitkin, 1991). Lavine and Wolpert (1995) even argue that there might never be any satisfactory solution to that problem. Her real concern is that she might end up with an unrealistically high Bayes Factor in favour of the null hypothesis represented by the transition probabilities in Table B.1. If she chose an improper prior over the alternative parameter space, that Factor could actually be infinite. But what if she chose Table B.1 supplemented by Tables B.2-4 as her null hypothesis instead? Could she not just as easily get an unrealistically high Bayes Factor in favour of that? And wasn't that hypothesis far more probable than Table B.1 alone? She decides that in her actual situation she doesn't have to worry; the set \( \{ T_u^E, T_u^N, T_u^W \} \) is indicated quite clearly enough. She chooses accordingly.

6. CONCLUDING REMARKS

Stone's stated intention with his Flatland puzzle was to demonstrate that improper priors, and in particular uniform improper priors, could lead to
strong inconsistencies. He saw the strong inconsistency in this instance as lying in the contrast between his Bayesian Statistician's subjective probability that he would win (3/4), and his coverage probability of winning (1/4).

We have found here, however, that merely using a uniform prior over all intersections in Flatland does not produce strong inconsistencies, even if Flatland is infinite in extent and the prior is diffuse. Any remaining inconsistencies cast doubt on the Classical approach rather than the Bayesian. The fact that Step 120 was a left turn is a very important piece of information in the empirical Bayesian analysis, but if the Classical approach is to obtain results that are similar in order of magnitude for the posterior odds on \( T_e \), this fact must be studiously ignored.

Barbara has it easy in this scenario. Had the treasure layers travelled anticlockwise around the loop in the north-west corner, she would not have been able to avoid the test of a precise null hypothesis against a composite alternative, the prospect of which so dismayed her at the end of Section 5. She is made to confront this more difficult problem in Part II.
II. The Anticlockwise Scenario

If the treasure layers traversed the loop in the north-west corner of Stone's diagram anticlockwise, the least probable location for the treasure cannot be determined without testing a precise null hypothesis against a composite and possibly diffuse alternative. This has long been a serious problem area for Bayesian inference, but is particularly so here, one observation being so different from the rest. A new general and automatic method is suggested for dealing with this problem, and the same choice of treasure locations is found to be optimal for both scenarios.

Previous approaches to the above problem have typically retained the traditional value unity for the prior odds ratio and then sought to specify plausible departures from the ideally diffuse alternative prior. For some of these, twice the logarithm of the Bayes Factor can be made asymptotically equal to the Bayes Information Criterion (BIC). Here, the diffuse alternative prior is retained and the prior odds ratio is adjusted so as to make twice the logarithm of the posterior odds equal to the BIC, up to an arbitrary scalar of the order of unity. Reasons are given for choosing this scalar to be $\sqrt{2}$. The procedure can also be generalized for use with non-diffuse alternative priors.

KEY WORDS: Bayes factor; Bayes information criterion; Empirical Bayes; Hypothesis testing; Noninformative prior; Posterior odds.
1. THE CHALLENGE OF THE ANTICLOCKWISE SCENARIO

It was shown in Part I that an empirical Bayes analysis of Stone's (1976) treasure hunt in Flatland could result in a solution different from and markedly preferable to the solution provided by his Classical Statistician, and indeed to any solution dependent only on the known and objective probabilities provided by the final random step. The particular demonstration provided was, however, contingent on the assumption that the treasure layers had travelled clockwise around the loop in the north-west corner of Stone's diagram. This permitted an important simplification to be made in the analysis.

We now consider what happens if Barbara, while exploring the path, discovers instead that the treasure layers travelled *anticlockwise* around the loop. (The notation and tables to be used here are the same as those employed in Part I.) The immediately consequential differences are that Steps 26-32 inclusive are now *NNWSS EE* (rather than *WNNES SE*) and *FFLLF IF* (instead of *LRFRR FL*). The marginal numbers of steps in the four cardinal directions remain unchanged, so Table B.1 is not affected, but the other relevant marginal distribution becomes *R 38, F 38, L 42*. Chi-square with two degrees of freedom is 0.34 for the CR assumption (equal proportions) and 0.49 for the 18:19:18 proportions implied by Table B.1, so the observed marginals do not conflict with either hypothesis. It is only when the numbers of runs of consecutive *R*s, *F*s and *L*s are considered that any important differences show up. For runs of two the counts are *RR 10, FF 11* and *LL 8*, total 29.
This is still not significantly different from 25.2, ie from 2/3 of the 38.4 expected from Table B.1. The crucial difference is that there is now an FFF sequence (Steps 25-27). This single occurrence is enough to make the alternative hypothesis constructed for the Clockwise Scenario (Table B.1 supplemented by Tables B.2-4) impossible—its likelihood is zero.

Sequences of three remain far fewer than would occur by chance. On an unconditional basis, as noted in Part I, Table B.1 implies that the expected number of runs of three is 12.9. On a conditional basis, summing over the 29 possible transitions from sequences of two to sequences of three, it is reduced to 9.7, but the single incidence of "success" remains significantly lower than the expectation by any reasonable criterion. However, all plausible alternatives are composite, and the best way to construct a reference prior for testing a precise null hypothesis against a composite alternative is a contentious issue.

The exact specification of the null hypothesis is also rather complicated, and any plausible specification of the alternative would be even more so. Barbara decides to consider first the far simpler but still reasonably comparable situation where 29 binomial trials have been conducted, resulting in one success and 28 failures. The null hypothesis is that the binomial parameter, $P$, is precisely 1/3. The alternative hypothesis is that $P$ is unknown and requires estimation. She uses this approximation as a preliminary guide to what might and might not be a helpful course to pursue in her actual situation.
2. THE REFERENCE PRIOR PROBLEM

The important statistic emerging from a Bayesian hypothesis test is the Bayes Factor, which is the ratio of the expected likelihoods for the two hypotheses over the null and alternative priors respectively. For a precise hypothesis, such as the null on this occasion \( (H_0; P = 1/3) \) the expected and actual likelihoods coincide. The Bayes Factor may be defined with the likelihood for the null hypothesis in either the numerator or the denominator. The former convention will be used for the discussion which follows.

There is an established consensus that a reference prior for a Bayesian hypothesis test should ascribe a prior probability of 1/2 to each of the two hypotheses. There are three distributions over the interval \([0,1]\) in the alternative parameter space that have some claim to be considered "neutral". The first is the uniform prior and the second is the Jeffreys prior, \( P^{-1/2}(1-P)^{-1/2} / \pi \). (Here \( \pi \) is not a probability but the number 3.14159...).

The third is the improper distribution proportional to \( P^{-1}(1-P)^{-1} \). For Barbara's binomial approximation the Bayes Factor is 0.0034029 using the uniform prior and 0.0021375 using the Jeffreys prior, but infinity using the improper distribution. Barbara considers that this infinite value is enough to invalidate the entire approach, and is determined to find one that is more intuitively plausible.
2.1 The Normal Distribution

She decides to start with the even simpler case of the normal distribution with known variance and unknown mean. There also the consensus is held that the two hypotheses should be regarded as equally probable before any observations are taken, and the same kind of paradoxical behaviour can be observed with the improper uniform prior over the real line. To quote from Lavine and Wolpert's (1995) comment:

"Before asking whether [O'Hagan's] solution is sensible we want to ask first whether any solution is sensible.

"Improper priors are often used in the hope that their posteriors approximate well the posterior that would have resulted from any well-thought-out prior. We typically reason that 'my prior is flat compared with the likelihood', 'there is much more information in the data than in the prior' and therefore 'my posterior is well approximated by the posterior from a convenient improper prior'. Then we adopt the improper prior and invest our effort more productively in other aspects of the analysis.

"But approximation of the prior is not the same as approximation of the Bayes Factor. When O'Hagan considers an improper prior without saying which proper prior's posteriors he is hoping to approximate then we should bear in mind all proper priors with
posteriors similar to that from the improper prior. If these priors were to yield roughly similar Bayes Factors, then it would be reasonable to associate a Bayes Factor (or a small range of Bayes Factors) with the improper prior ....  If, however, there are priors that yield posterior distributions similar to that from the improper prior, but that yield vastly different Bayes Factors, then the specification of a Bayes Factor for the improper prior is problematic at best."

Lavine and Wolpert then present a range of proper uniform priors that yield roughly the same posterior distribution as O'Hagan's improper prior (uniform on the real line) but with Bayes Factors that range from around $10^{-10}$ up to infinity. They continue:

"We conclude that there is no uniquely acceptable Bayes Factor, even approximately ....  When comparing models the investigator simply cannot shirk the responsibility of specifying the prior distribution (and hence the alternative hypothesis) in more detail."

By implication, Lavine and Wolpert require a subjective specification, and hence a Bayes Factor that might vary considerably from experimenter to experimenter. From their own example it is clear that it can in fact vary quite wildly. Barbara accepts their point that the current procedure is fatally flawed, but refuses to accept their subjectivist prescription, which she regards as hardly any different in practice from allowing a licence for anyone to choose whatever answer they want. She considers instead the various attempts that
have been made to circumvent the inescapable fact that improper priors permit (and to all intents ensure) infinite Bayes Factors.

So far as the normal distribution is concerned, she distinguishes four basic approaches that have been suggested for modification of the uniform prior over the real line. All four imply, in one fashion or another, the prior existence of observations before the experiment begins. Two can be characterized as conservative, in that they limit this imaginary prior information to the equivalent of a single observation or less. The other two are radical or "replicated", requiring as many imaginary prior observations as there are real ones. On another axis of classification, two of them centre these imaginary observations at the null hypothesis value, while the other two centre them at or near the mean of the actual observations. Typical examples of these approaches can be found in the following references.


Replicated, centred on null. Smith and Spiegelhalter (1980).


It is convenient to consider these four approaches in reverse order.
Orthodox Bayesians rejected Aitkin's use of the expression, "Posterior Bayes Factor", from its inception. It was felt to confuse prior with posterior. Fearn (1991) commented:

"The author, as he points out, has a perfect right to calculate the posterior mean of any quantity he wishes [in this case the likelihood]—what he cannot do is to interpret the result as a Bayes Factor, except in so far as it corresponds to one very special and informative prior."

Typical of the less tolerant comments was this one by Cuzick (1991):

"The most problematic aspect of Professor Aitkin's proposal is to use the same data twice for making a single inference. ... A more defensible approach would be to split the data into two samples, one for estimating the prior and the second for computing the expected likelihood."

This notion of using part of the data as a "training sample" could be reckoned as a fifth approach, except that its arbitrary nature makes it so obviously suboptimal. The idea was nevertheless the initial motivation behind O'Hagan's suggestion of "Fractional Bayes Factors". That author minimized the loss of information by taking the training sample to be as small as possible (a single observation) and then optimized its position by taking it to be at the mean of the whole sample. A.F.M. Smith (1991), however, in seconding the vote of thanks to O'Hagan, pointed out that the procedure was not strictly coherent. This point was also picked up by several other discussants. (The mean of the observations is not known until the entire set has been observed.)
The "Intrinsic Bayes Factor" approach used by Berger and Pericchi (1996) is very similar to O'Hagan's in practice, but rather than using a fractional part of the entire likelihood it aims for a minimum training sample from the start and then examines the properties of the arithmetic and geometric means of all available unit training samples, and also those of the expected values of these two means, evaluated at the position of the maximum likelihood estimator (MLE). Again, none of these is strictly coherent.

The replicated approach suggested by Smith and Spiegelhalter (1980) is coherent only if the sample size is chosen beforehand and adhered to strictly. If, however, it proves impossible to obtain the size originally decided on, or if that size is later increased, the prior has to be changed accordingly, resulting in incoherence. In any case, the resulting prior is highly informative, and therefore not the kind that Barbara needs for her problem.

This leaves the single (or fractional) observation, centred on the null, as the only coherent contender. Jeffreys (1961) advocated introducing a single observation, Cauchy distributed. Berger (1995) used, for illustrative purposes, an observation that was normally distributed but with twice the variance of the actual observations. Each of these could be regarded heuristically as equivalent to half an observation (Berger and Pericchi 1996). Given a sufficiently large number of observations, these two procedures produce plausible Bayes Factors, as indeed do the Fractional and Intrinsic Bayes Factor approaches. Kass and Wasserman (1995) show that they provide, asymptotically, Bayes Factors equal to the exponents of the Schwarz Criteria,
S, (Schwarz 1978). [They note that the Schwarz Criterion is "also known as
the Bayes Information Criterion or BIC, though in the usual definition 2S =
BIC". (The two are equivalent in practice.)]

But asymptotic properties of this kind are no immediate help with Barbara's
binomial problem. Although she has a sample of 29, she has only one
"success". Even an imaginary one-third of a "successful" observation that
was combined with an imaginary two-thirds of an "unsuccessful" one would
bias her prior unduly towards the null hypothesis value of 1/3. (The
difference between $P^{-1}(1 - P)^{-1}$ and $P^{-1/2}(1 - P)^{-1/2}/\pi$, and likewise the
difference between $P^{-1/2}(1 - P)^{-1/2}/\pi$ and the uniform, can each be viewed as
equivalent to the introduction of two half-observations, one successful and one
unsuccessful. When—as in the case under study—the proportion of
observed successes is smaller than that expected under the null hypothesis, the
uniform alternative prior can be expected to favour the null most and the
improper prior favour it least.) She considers the possibility of using an
imaginary observation that is $r/n$ parts "successful" and $(n-r)/n$ parts
"unsuccessful", but rejects even this as not strictly coherent.

She returns to the normal distribution problem. If the alternative prior is
diffuse, the Bayes Factor is necessarily infinite, regardless of how small the
sample size might be or how extreme the observations are. This is because
almost all the alternative prior probability is infinitely remote from any finite
value. However extreme the actual observations, the likelihood will be
infinitely larger at the null value than at $\pm\infty$. But this concentration of the
alternative prior probability at the extremes is purely a convenient fiction.
Nobody really believes almost all the alternative prior probability to be infinitely remote from the null. The solution might lie, she conjectures, in balancing one convenient fiction against another. The fact that any finite prior odds multiplied by an infinite Bayes Factor will result in an infinite posterior odds ratio provides her with a necessary clue. She will not aim for any finite Bayes Factor, nor even directly for any reference prior odds, but rather for a reference posterior odds given some standard situation. The fact that the amount of information in a unit observation already possesses some status as a suitable reference occurs to her. How would she feel if, starting from a state of ignorance or near-ignorance, she found her first observation to be somewhere in the vicinity of the null hypothesis value? Would she not then continue to feel more or less indifferent between the null and the alternative hypotheses?

She assesses the consequences. The most useful reference prior she can imagine has almost all the prior probability given to the alternative hypothesis, but improperly and uniformly distributed over the real line. A positive but infinitesimal amount of probability, smaller than any real number, is located at the null hypothesis value. This prior is the limiting case of an infinite sequence of proper uniform priors, all of which have much the same ratio of null hypothesis probability to the amount of alternative hypothesis probability contained within a given narrow range (say one observation's standard deviation) of the null hypothesis value. Moreover, all these more realistic priors have very similar posterior odds ratios for any given realistic set of observations including, most importantly, the case of a single observation in the close neighbourhood of the null hypothesis value.
To formalize this idea, she specifies the incoming observations to be $N(\mu, \sigma^2)$ with $\sigma^2$ known and $\mu$ unknown. The null hypothesis is $H_0: \mu = \mu_0$ and the alternative, $H_1$, that $\mu$ is distributed uniformly over $(\mu_0 - C, \mu_0 + C)$ where $C \gg \sigma$. Prior probability $\pi_0$ is associated with $H_0$ and $\pi_1 = 1 - \pi_0$ with $H_1$. Provided the true value of $\mu$ lies in the interval $(\mu_0 - C + 5\sigma, \mu_0 + C - 5\sigma)$ it is nearly certain that the posterior odds will be closely approximated by $(\pi_0 / \pi_1) 2C / \sqrt{2\pi\sigma^2} \exp\{-n(\bar{y} - \mu_0)^2 / 2\sigma^2\}$, where $\bar{y}$ is the mean of the $n$ observations $\mathbf{y} = (y_1, y_2, \ldots, y_n)$. Barbara’s choice of reference criterion requires that when $n = 1$ and $y_i = \bar{y} = \mu_0$ the posterior odds should take a value reasonably close to unity. Denoting this value by $k$, she obtains

$$\frac{\pi_0}{\pi_1} = k / \left[2C / \sqrt{2\pi\sigma^2}\right] \quad \text{or} \quad \pi_0 = \frac{k\sigma\sqrt{2\pi} / 2C}{1 + \left[k\sigma\sqrt{2\pi} / 2C\right]}.$$

This implies that $\lim_{C \to \infty} C\pi_0 = \sqrt{\pi} / 2k\sigma$, which may be interpreted as requiring that the amount of prior probability associated with $H_0$ should be equal to the amount of alternative prior probability contained in the interval $(\mu_0 - \sqrt{\pi}/2k\sigma, \mu_0 + \sqrt{\pi}/2k\sigma)$. In this limiting case the posterior odds are $k\sqrt{n} \exp\{-n(\bar{y} - \mu_0)^2 / 2\sigma^2\}$ and, if in addition $n = 1$, they attain their maximum value of $k$ when $y_i = \bar{y} = \mu_0$.

The most obvious candidate value for $k$ is unity. The logarithm of the posterior odds would then be exactly equal to the univariate version of the
Schwarz Criterion, S. However the choice $k = 1$ would imply that after a single observation the most favourable possible attitude towards the null hypothesis would be indifference. This is not in accord with Barbara's ideas as to what a reference prior should be capable of achieving. She prefers the option $k = \sqrt{2}$, suggested by the fact that the average height of the normal distribution is $1/\sqrt{2}$ of its maximum height. [This is the reason why Aitkin (1991) incorporated a factor $\sqrt{2}$ into his definition of the Posterior Bayes Factor (PBF).] Barbara's posterior odds are therefore

$$\sqrt{2n} \exp\{-n(\bar{y} - \mu_0)^2 / 2\sigma^2\},$$

the factor $\sqrt{2}$ coming effectively from the PBF and the remainder of the expression reflecting the Schwarz Criterion.

Given $k = \sqrt{2}$, indifference occurs when $y_i = \mu_0 \pm \sigma \sqrt{\ln 2}$. The numerical value of $\sqrt{\ln 2}$ is very nearly $5/6$ and the proportion of the standard normal distribution within $\sqrt{\ln 2}$ either side of the mean is close to 80 per cent. So if $H_0$ is true, there are some four chances in five that the first observation will indicate mild support for that hypothesis. Further, since $k = \sqrt{2}$ implies

$$\lim_{C \rightarrow \infty} C \pi_0 = \sqrt{\pi} \sigma,$$

the infinitesimal amount of prior probability associated with $H_0$ is equal to the infinitesimal amount of alternative prior probability contained in the interval $(\mu_0 - \sqrt{\pi} \sigma, \mu_0 + \sqrt{\pi} \sigma)$ or, equivalently, in an interval of length $2\sqrt{\pi} \sigma$ anywhere on the real line. $2\sqrt{\pi} \sigma$ is approximately 3.545, and the interval $(\mu - \sqrt{\pi} \sigma, \mu + \sqrt{\pi} \sigma)$ contains over 92 per cent of $N(\mu, \sigma^2)$.

2.2 The Binomial Distribution

Barbara now translates her reference prior criterion into more general terms. In the context of her binomial problem, although she cannot imagine it she can
at least describe mathematically a first observation in exact agreement with the null hypothesis. In her case it would have likelihood \((1/3)^{1/3}(2/3)^{2/3}\). But she prefers to consider the normal limit of the binomial distribution as \(n\) tends to infinity. If the prior is \(H_0: P = P_0\) and there are \(r\) successes among the \(n\) observations where \(n\) is large, the distribution of \(\bar{y} = r / n\) should be close to \(N(P_0, P_0 Q_0 / n)\), where \(Q_0 = 1 - P_0\), and the posterior odds in favour of \(H_0\) should be \(\sqrt{2n}\) when \(r = P_0 n\).

Using the conjugate prior

\[
\frac{\Gamma(2\alpha)}{[\Gamma(\alpha)]^2} P_0^{\alpha - 1} (1 - P_0)^{\alpha - 1}
\]

and given an experiment with \(r\) successes and \((n-r)\) failures, the posterior odds in favour of \(H_0\) are

\[
\frac{\text{Post Pr}(H_0)}{\text{Post Pr}(H_1)} = \frac{(\pi_0 / \pi_1) P_0^\alpha Q_0^{n-r} [\Gamma(\alpha)]^2 / \Gamma(2\alpha)}{\Gamma(r + \alpha) \Gamma(n - r + \alpha) / \Gamma(n + 2\alpha)}
\]

Applying Stirling's formula to the Gamma functions in the denominator,

\[
\frac{\text{Post Pr}(H_0)}{\text{Post Pr}(H_1)} = \frac{(\pi_0 / \pi_1) P_0^\alpha Q_0^{n-r} [\Gamma(\alpha)]^2 / \Gamma(2\alpha)}{\sqrt{2\pi(n + 2\alpha - 1)} 2\pi(n - r + \alpha - 1)} \frac{(r + \alpha - 1)^{n+r+\alpha-1} \exp(n + 2\alpha - 1)}{(n+2\alpha-1)^{n+2\alpha-1} \exp(n - r + \alpha - 1)}
\]
\[
\pi_0 \cdot \frac{\Gamma(\alpha)^2}{\pi_1} \cdot \frac{P_0 Q_0^{\alpha-r}}{\Gamma(2\alpha)} \cdot \frac{\left[\frac{2\pi r(n-r)}{n^{r+\alpha-1}} \exp(\alpha-1)(n-r)^{n-r+\alpha-1} \exp(\alpha-1) \exp(n+2\alpha-1)\right]}{\sqrt{n^{n+2\alpha-1} \exp(2\alpha-1) \exp(r+\alpha-1) \exp(n-r+\alpha-1)}}
\]

\[
\pi_0 \cdot \frac{P_0 Q_0^{\alpha-r}}{\pi_1} \cdot \frac{\Gamma(\alpha)^2}{\Gamma(2\alpha)} \cdot \frac{\sqrt{n}}{2\pi(pq)^{3\alpha-1}}.
\]

where \( p = r/n \) and \( q = 1-p \).

Barbara's generalized criterion for a unidimensional reference prior requires that when \( n \) is large and the observations are such that the posterior odds ratio attains its maximum possible value conditional on \( n \), its leading term should be \( \sqrt{2n} \); this is so here when \( p \equiv P_0 \), so her prior odds are

\[
\frac{\pi_0}{\pi_1} = \frac{\Gamma(2\alpha)}{(\Gamma(\alpha))^2} \cdot \frac{2\pi(P_0 Q_0)^{2\alpha-1}}{\sqrt{2n}} = 2\sqrt{\pi(P_0 Q_0)^{2\alpha-1} \Gamma(2\alpha)} / (\Gamma(\alpha))^2.
\]

For \( \alpha = 1 \), \( \pi_0 / \pi_1 = 2\sqrt{\pi P_0 Q_0} \). For \( P_0 = 1/3 \), \( \pi_0 / \pi_1 = 2\sqrt{2\pi} / 3 \equiv 1.671 \) and

\[
\frac{\text{Post Pr}(H_0)}{\text{Post Pr}(H_1)} = \frac{(2\sqrt{2\pi}/3)(1/3)(2/3)^{2\alpha}}{\Gamma(2) \Gamma(29) / \Gamma(31)} \equiv 0.005686 \equiv 1/175.9.
\]

For the Jeffreys prior, \( \alpha = 1/2 \) and \( \pi_0 / \pi_1 = 2/\sqrt{\pi} \), the factor in \( P_0 Q_0 \) cancelling out, and \( \pi_0 / \pi_1 \equiv 1.128 \) for all permissible values of \( P_0 \). So for any precise value of \( P_0 \),

\[
\frac{\text{Post Pr}(H_0)}{\text{Post Pr}(H_1)} = \frac{(2/\sqrt{\pi})(1/3)(2/3)^{2\alpha}}{\Gamma(1.5) \Gamma(28.5) / \Gamma(30)} \equiv 0.002412 \equiv 1/414.7.
\]
When $\alpha = \varepsilon$ (where $\varepsilon < 1$), $\pi_0 / \pi_i = \varepsilon \sqrt{\pi_i / (P_0 Q_0)}$. For $P_0 = 1/3$, $\pi_0 / \pi_i \equiv 3.760 \varepsilon$ and

$$
\frac{Post \ Pr(H_0)}{Post \ Pr(H_1)} = \frac{\varepsilon^{3\sqrt{\pi/2}}(1/3)(2/3)^{28}(2/\varepsilon)}{\Gamma(1) \Gamma(28) / \Gamma(29)} \equiv 0.000824 \equiv 1/1214.2.
$$

As anticipated earlier in this Subsection, the uniform alternative prior favours the null hypothesis most and the improper prior favours it least. Since the $P_0 Q_0$ factor disappears when the Jeffreys prior is used, Barbara considers this to have the best claim to be used as a reference or standard prior. To be on the safe side, however, she plans to use the uniform prior in her sensitivity tests.

She next considers whether her own subjective prior beliefs are compatible with those she obtained using these reference priors. She looks to make sure that she is not being influenced unduly by the fact that she devised these reference priors herself (after all, there is a substantial treasure at stake) but cannot see any reason for departing from her own criterion on this occasion. The prior odds for the Jeffreys prior in particular look very plausible. She decides that no subjective adjustment to these prior odds would have been necessary had her problem been strictly binomial. But of course it was not. She must now meet the "real life" demands of her Flatland treasure hunt!
There are two relevant hypotheses for the Anticlockwise Scenario. The simple null is defined by the transition probabilities in Table B.1. (It is "simple" only in the technical sense that all the probabilities are known exactly and there is no parameter left to estimate). The composite alternative also follows Table B.1 wherever the previous two steps are of different kinds (eg an L and an R, or an R and an F, but not two L's, two R's or two F's in succession). It must, however, have rules to define what transition probabilities are to be used for the 29 steps that immediately follow the events of the types RR, FF and LL.

Barbara has yet to specify what these rules should be. (As in Section 1, only these steps need be considered when constructing a likelihood function, as the two hypotheses have the same likelihoods for the remainder.)

Her simplest plausible specification for these rules gives weight $\lambda$ to Table B.1 and weight $1 - \lambda$ to the other relevant Table. For any given value of $\lambda$ the likelihood of the 29 relevant independent observations is then

$$
\text{Lik}(\lambda; x') = \frac{1}{3} \lambda \left(1 - \frac{3}{10} \lambda\right)^6 \left(1 - \frac{2}{9} \lambda\right)^6 \left(1 - \frac{3}{8} \lambda\right)^3 \left(1 - \frac{3}{8} \lambda\right)^3
$$

the null hypothesis being $H_0: \lambda = 1$. Equivalently, writing $\theta$ for $\lambda / 3$, the likelihood is

$$
\text{Lik}(\theta; x') = \theta \left(1 - \frac{9}{10} \theta\right)^6 \left(1 - \frac{12}{10} \theta\right)^6 \left(1 - \frac{6}{9} \theta\right)^7 \left(1 - \frac{12}{9} \theta\right)^3 \left(1 - \frac{9}{8} \theta\right)^3.
$$
and the null hypothesis is $H_0: \theta = 1/3$. Noting that for the binomial approximation the alternative hypothesis had its prior probability defined over the interval (0,1) for $P$, and that this interval now corresponds to (0,1) for $\theta$, she decides to keep things as comparable as possible and use the conjugate prior

$$f(\theta, \alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \theta^{a-1}(1-\theta)^{a-1}$$

over the same interval, again for the three cases $\alpha = 1$, $\alpha = 1/2$ and the limit as $\alpha$ approaches zero. She realises that $\text{Lik}(\theta; x')$ can take negative values in some parts of the interval (3/4,1), but guesses (correctly) that she can replace all likelihoods in that interval by zero without appreciable loss of precision.

She then makes the following three adjustments to the posterior odds that she used for the binomial approximation:

(i) in the expressions for $\pi_0 / \pi_1$, replacing the factor $(1/3)(2/3)$ or 0.2222 with the variance, under $H_0$, of the mean of the 29 relevant binomial variables, i.e.

$$\left\{ 6\left(\frac{3}{10}, \frac{7}{10}\right) + 3\left(\frac{4}{10}, \frac{6}{10}\right) + 6\left(\frac{2}{9}, \frac{7}{9}\right) + 8\left(\frac{3}{9}, \frac{6}{9}\right) + 3\left(\frac{4}{9}, \frac{5}{9}\right) + 3\left(\frac{3}{8}, \frac{5}{8}\right) \right\} / 29 = 0.2151$$

(ii) for the likelihood under $H_0$, replacing the factor $(1/3)(2/3)^8$ or $3.911 \times 10^{-6}$ with
\[ \text{Lik}(1/3; x') = \frac{1}{3} \left( \frac{7}{10} \right)^6 \left( \frac{6}{10} \right)^3 \left( \frac{7}{9} \right)^6 \left( \frac{5}{9} \right)^3 \left( \frac{5}{8} \right)^3 = \frac{7^{12}}{3^{21} 2^8 5^3} \equiv 4.594 \times 10^{-6} \]

and

(iii) for the mean likelihood under \( H_1 \), replacing the divisor

\[ \Gamma(r + \alpha) \Gamma(n - r + \alpha) / \Gamma(n + 2\alpha) \]

with \( \int_0^{3/4} \text{Lik}(\theta; x') f(\theta, \alpha) d\theta \).

The values of \( \Gamma(r + \alpha) \Gamma(n - r + \alpha) / \Gamma(n + 2\alpha) \) are 0.0011494 for \( \alpha = 1 \),
0.0057495 for \( \alpha = 1/2 \), and 0.035714 for the limit as \( \alpha \) approaches zero.

The corresponding values for the integral are 0.0011965, 0.0059319 and 0.036495 respectively.

The reference posterior odds are therefore,

a) for the case \( \alpha = 1 \),

\[ \frac{\text{Post Pr}(H_0)}{\text{Post Pr}(H_1)} \equiv 0.005686 \left( \frac{0.2151 \cdot 4.594 \times 10^{-6} \cdot 0.0011494}{0.2222 \cdot 3.911 \times 10^{-6} \cdot 0.0011965} \right) \equiv 0.006212 \equiv 1/161.0 \]

b) for the case \( \alpha = 1/2 \),

\[ \frac{\text{Post Pr}(H_0)}{\text{Post Pr}(H_1)} \equiv 0.002411 \left( \frac{0.2151 \cdot 4.594 \times 10^{-6} \cdot 0.0057495}{0.2222 \cdot 3.911 \times 10^{-6} \cdot 0.0059319} \right) \equiv 0.002657 \equiv 1/376.3 \]

c) for the limit as \( \alpha \) approaches zero,
\[
\frac{\text{Post Pr}(H_0)}{\text{Post Pr}(H_1)} = 0.000824 \left( \begin{array}{c}
0.2151 \quad 4.594 \times 10^{-6} \quad 0.035714 \\
0.2222 \quad 3.911 \times 10^{-6} \quad 0.036495 \\
\end{array} \right) \equiv 0.000917 \equiv 1/1090.6
\]

Nowhere is there any departure from the binomial approximation sufficient to cause any practical concern. The weight to be given to Table B.1 when considering the relative posterior probabilities for \( T_w^w \) and \( T_a^s \) is in each case well below one per cent. Given \( H_1 \), the weight to be given to Table B.1 is the posterior mean of \( \theta \) (under \( H_1 \)) which is given by \( \bar{\lambda} = \frac{\theta}{3} \) where

\[
\bar{\theta} = \int_0^{3/4} \theta \text{Lik}(\theta; x') f(\theta, \alpha) d\theta / \int_0^{3/4} \text{Lik}(\theta; x') f(\theta, \alpha) d\theta.
\]

For \( \alpha = 1 \), \( \bar{\theta} = 0.0657 \) and \( \bar{\lambda} = 0.197 \), for \( \alpha = 1/2 \), \( \bar{\theta} = 0.0509 \) and \( \bar{\lambda} = 0.153 \) and for the limiting case as \( \alpha \) approaches zero, \( \bar{\theta} = 0.0352 \) and \( \bar{\lambda} = 0.106 \). In none of these cases does the combination of the posterior probability of the null hypothesis and the value of \( \bar{\lambda} \) under the alternative bring the weight to be applied to Table B.1 anywhere near 16/21. Barbara therefore decides quite confidently that \( T_a^s \) is easily the least probable location for the treasure and makes the same choice in the Anticlockwise Scenario as she makes in the clockwise, i.e. \( \{ T_a^e, T_a^n, T_w^w \} \).

4. CONCLUDING REMARKS

The two scenarios result in the same optimal set of potential treasure locations, and the assumption of a uniform prior over all intersections is as innocuous in the anticlockwise one as in the clockwise. Despite this, a different kind of diffuse prior has been seen to present problems in the Anticlockwise Scenario. The test of a precise null hypothesis against a diffuse alternative results unequivocally in an infinite Bayes Factor. Lavine and Wolpert (1995)
have already been quoted in this context. Berger and Pericchi (1996) are equally uncompromising. After describing how Jeffreys (1961) chooses the properties for his proper (Cauchy) prior they continue:

"Although it is easy to object to having such choices 'imposed' on the analysis, it is crucial to keep in mind that there is no alternative (except subjective elicitation). Alternative default methods either themselves correspond to imposition of some (proper) default prior or, worse, end up not corresponding to any actual Bayesian analysis. This issue is important enough to deserve emphasis."

We could perhaps identify the "invisible but valuable treasure" that Barbara seems set to win in Section 3 with her recipe for circumventing this long-standing dilemma. She does not modify the diffuse alternative prior in any way. Instead she balances the unrealistically spread-out nature of this alternative by assigning a corresponding unrealistically small probability to the null. In doing so, she succeeds in gives the Schwarz criterion (and the BIC) a new interpretation, at least for the normal distribution. They are not only (twice) the logarithm of an asymptotic approximation to a possible Bayes Factor, but exactly (or exactly twice) the logarithm of a feasible posterior odds ratio. (The prior probability on $H_0$ required to yield the Schwarz criterion as the logarithm of the posterior odds ratio is actually $1/\sqrt{2}$ of Barbara's reference prior probability.)

Further relationships, linking this odds ratio to other relevant statistics including Aitkin's (1991) Posterior Bayes Factor and the corresponding
Classical or Frequentist hypothesis test statistic, together with the extension of Barbara's criterion to \( v \) dimensions, are considered in Part III.

**ADDENDUM**

At a late stage in the preparation of this thesis, I received a private communication from Prof. D.V. Lindley, drawing attention to the reasons given by Jeffreys (1961, §5.2) for choosing a Cauchy prior. The relevant portion reads:

"There he says that the result of one observation should not affect his opinion on \( H_0 \) because any departure from zero could be attributed to \( \sigma \). (He is taking the case where \( \sigma \) is unknown.) This is satisfied if the prior is an even function with integral one. He next considers what he might think if he were to have two observations which agreed and were not zero. This should convince him that both \( \sigma = 0 \) and \( H_1 \) is true. This is satisfied if the prior is Cauchy about zero. Notice he is not saying that his prior is equivalent to imaginary observations, nor is he using part of the data to obtain a prior. He is saying that inference is obvious for some special cases \((n = 1, 2)\) and that the result for general \( n \) should include these. Two puzzles remain for me. I have never understood his statement on 1.4 of §5.2 'From considerations of similarity it [the prior on the alternative] must depend on \( \sigma \), since there is nothing in the problem except \( \sigma \) to give a scale for \( \lambda[\mu] \).' In §5.21 he continues to use the Cauchy in the case where \( \sigma \) is known (your case) and the scale exists. Apart from that, I think Jeffreys's argument is brilliant and a fine example of using coherence to establish a prior. Essentially, you know the probability here \((n = 1, 2)\), so we can deduce it for general \( n \) (including zero)."

It seems that in working backwards from \( n = 1 \) to \( n = 0 \) I was unwittingly following a precedent set by Jeffreys.
III. Properties of The Reference Posterior Odds

A new procedure was introduced in Part II for testing a precise null hypothesis against a composite alternative. It used posterior odds rather than Bayes Factors and tailored the prior odds to take account of how thinly the alternative prior was spread over the available parameter space. The properties of the resulting Reference Posterior Odds ratio are examined here. It is shown to be closely related to the ratio of the Bayes Factor to its repeated sampling expectation, to Aitkin's (1991) Posterior Bayes Factor and also to the density or distribution of the corresponding Frequentist test statistic, differing from all three principally by a factor \( \sqrt{n} \). This factor reflects the rate at which the Bayes Factor tends to increase with the size of the experiment when the null hypothesis is true. There is also an even closer relationship between the Reference Posterior Odds ratio and the Schwarz Criterion/BIC, the former being just slightly more parsimonious than the latter.

Methods that rely on the Bayes Factor alone as a measure of the empirical evidence suffer from the problem of its extreme sensitivity to small changes in the specification of the alternative prior. The Reference Posterior Odds ratio is not subject to this problem.

Examples are given as to how the method can be used to test for a particular value of a normal mean, a normal variance and a binomial parameter. For this last, three special cases are considered; the uniform prior, the Jeffreys prior and the improper prior proportional to \( 1/[P(1 - P)] \). The three Reference Posterior Odds ratios are shown to be close to each other in value, except when there are only a few "successes" or only a few "failures" among the observations.
The close relationship between the Reference Prior Odds ratio and the Frequentist test statistic enables a Bayesian interpretation to be made of the Frequentist test procedure. It is coherent, but only for an experiment of fixed size. This prompts a suggestion that the two approaches have more to offer each other than is commonly supposed.

KEY WORDS: Bayes factor; Bayes-Frequentist interface; Bayes information criterion; Noninformative prior; Posterior Bayes factor.

1. INTRODUCTION

Defining an appropriate reference prior for testing a precise null hypothesis against a composite—and often ideally improper—alternative has long been a vexatious issue for Reference Bayesian statisticians. The basic problem is that for any improper alternative prior [such as the uniform over the real line for the mean of a normal distribution, \( \mu \), or one proportional to \( P^{-1}(1 - P)^{-1} \) for the binomial parameter, \( P \)] the resulting Bayes Factor is automatically infinite—no matter how small the sample or how extreme the observations. This problem is usually handled by specifying a minor departure from the ideally improper prior, small enough not to affect the posterior distribution materially for a sample of moderate size but large enough to allow the Bayes Factor to take plausible values. However, as pointed out forcibly by Lavine and Wolpert (1995), the very fact that small changes to the posterior distribution can be accompanied by spectacular changes in the Bayes Factor makes it unlikely that any meaningful reference hypothesis test can ever be specified using this approach.
In Part II, this hypersensitive behaviour of the Bayes Factor was circumvented by concentrating on the posterior odds ratio instead. The ideal alternative prior was retained, even if improper, and the prior odds were chosen so as to ensure that when the sample size, $n$, was large and the observations such that the posterior odds ratio attained its maximum possible value conditional on $n$, the leading term of that ratio should be equal to $k\sqrt{n}$, where $k$ was an arbitrary scalar. For normally distributed observations, the choice $k = 1$ ensured that the reference value for the posterior odds was exactly equal to the exponent of the Schwarz criterion (i.e., to the exponent of half the Bayes Information Criterion or BIC). Reasons for preferring $k = \sqrt{2}$ were advanced, but the choice remained essentially arbitrary.

Some properties of the resulting Reference Posterior Odds ratio (RPO or $G_r$) are described in the following Sections of Part III. As a preliminary, the ordinary Bayes Factor, its repeated sampling expectation, and the ratio of the two (the "Realization Factor", $R$) are discussed in Section 2. The RPO itself is described in Section 3, and some examples are of it are given. In Section 4, Aitkin's Posterior Bayes Factor (PBF or $A$) is shown to be closely related to both $R$ and $G_r$. In Section 5, all three of these are shown to be related to the Classical or Frequentist test statistic as well. Implications for the choice of information criterion are considered in Section 6. The closing remarks in Section 7 are chiefly directed towards illuminating the interface between the Bayesian and the Classical (or Frequentist) approaches to statistical inference.
2. THE BAYES FACTOR AND ITS EXPECTATION

The ordinary or conventional Bayes Factor is defined as the ratio of the mean likelihood of the parameter $\theta$ over the null hypothesis, $H_0$, to its mean likelihood over the alternative hypothesis, $H_1$. If the prior distribution under $H_i$ is $f_i(\theta)$, $i = 0,1$, the Bayes Factor may be written,

$$B = \frac{\int_{\theta \in H_0} \text{Lik}(\theta; y) f_0(\theta) d\theta}{\int_{\theta \in H_1} \text{Lik}(\theta; y) f_1(\theta) d\theta}.$$  

$y$ being the observation vector and $\int_{\theta \in H_i} f_i(\theta) d\theta$ being unity by definition. In the important special case where the null hypothesis is precise, i.e. of the form $H_0: \theta = \theta_0$, this expression becomes

$$B = \frac{\text{Lik}(\theta_0; y)}{\int_{\theta \in H_1} \text{Lik}(\theta; y) f_1(\theta) d\theta}.$$  

So defined, it is unequivocally a measure of the empirical evidence supporting $H_0$ over $H_1$, and this has led to it being regarded as the appropriate test statistic for the choice between these two hypotheses. This is prima facie the case in situations where $H_0$ and $H_1$ are genuinely practical alternatives and meaningful subjective prior probabilities can be associated with them, but, as pointed out by Aitkin (1991, p140, in reply to Lindley 1991) the use of $B$ alone to choose among four models (or, equivalently, hypotheses) can lead to a serious paradox. Denoting these models by $M_{11}$, $M_{12}$, $M_{21}$, and $M_{22}$, a
situation can arise where $M_{11}$ is favoured over $M_{12}$ and also $M_{21}$ over $M_{22}$, but $M_{12} \cup M_{22}$ is favoured over $M_{11} \cup M_{21}$.

Quite apart from this "Four Model Paradox," reliance on $B$ alone can present a particularly intractable problem for the Reference Bayesian (RB) approach (described in Bernard 1996) even when the choice is between two hypotheses only. As mentioned in Section 1, if $H_0$ is precise and $H_i$ is improper, the Bayes Factor is automatically infinite; hence if any finite prior probability, $\pi_0$, however small, is associated with $H_0$, the corresponding posterior probability is one, regardless of the experimental outcome. (Conventionally, $\pi_0$ and its complement, $\pi_1$, are each set at one half.) This second paradox can only be fully resolved by recognizing that such a prior distribution is not neutral but, in effect, biased overwhelmingly towards the precise $H_0$. An appropriate adjustment, making the prior more genuinely neutral, will be described in Section 3. In the meantime, however, it is relevant to note that the ratio of $B$ to its repeated sampling expectation under $H_0$ is stable against small changes to the distribution used to specify $H_i$.

Denoting this ratio by $R$, we have by definition that

$$R = \frac{B}{E_0(B)} = \frac{\text{Lik}(\theta_0; y) / E_1[\text{Lik}(\theta; y)]}{\int [\text{Lik}(\theta_0; y) / E_1[\text{Lik}(\theta; y)]] f_0(y) dy}$$

where $E_i$ denotes the expectation under $H_i$, $i = 0, 1$, $E_0(.)$ denotes the repeated sampling expectation under $H_0$ (calculated on the assumption that
the experiment is of fixed size) and \( f_0(y) \) is the sampling density of \( y \), also under \( H_0 \).

\( R \) is not the kind of statistic typically regarded as useful by Bayesian statisticians but, being directly proportional to \( B \), it is at least arguably an alternative measure of the empirical evidence favouring \( H_0 \) over \( H_1 \). We refer to it here as "the Realization Factor" because it measures the extent to which the repeated sampling expectation of \( B \) under \( H_0 \) is realized in the experimental situation. It will be demonstrated in Section 5 that there is a close relationship between \( R \) and the Classical or Frequentist test statistic.

3. THE REFERENCE POSTERIOR ODDS

Another alternative to the Bayes Factor is the Posterior Odds ratio. From a Bayesian standpoint, it is preferable to the Realization Factor, since it makes no use of repeated sampling (and hence no use of events that have not been observed). It is also attractive because, provided only that the assumed prior odds obey the laws of probability, the use of posterior odds ratios rather than Bayes Factors automatically eliminates the Four Model Paradox mentioned in Section 2. They can moreover be used—though less straightforwardly—to avoid the other paradox described in that Section.

A reference version of the posterior odds ratio (RPO) was used instead of \( B \) in the procedure ascribed to "Barbara" in Part II. The prior odds ratio was no longer given its conventional value of unity but was chosen rather to ensure
that twice the logarithm of the posterior odds was closely related to the Bayes Information Criterion (BIC). A condition of equivalence was met exactly (up to an arbitrary constant $2\ln k$) when the incoming observations were normal and i.i.d., and therefore also asymptotically wherever the Central Limit Theorem held. Where the alternative prior was ideally improper (as for the normal mean) the RPO was constrained to remain finite by considering a limiting process. As the alternative prior was allowed to become more diffuse, $B$ became larger but the prior odds were reduced correspondingly and the RPO remained constant. In the limit, where the alternative prior became improper, $B$ became infinite but the prior odds became correspondingly infinitesimal, and the RPO still remained unchanged.

This was actually a revival of a much earlier idea. M. S. Bartlett, on reading Lindley (1957), had written to the author pointing out that when the prior distribution describing $H_1$ was improper, $B$ became infinite and the "silly answer" ensued that the posterior probability of $H_0$ would be unity for any set of observations. Lindley's counter-suggestion was to make "the prior odds in favour of the null hypothesis against any unit interval of the alternative values [equal to a constant]" (Bartlett 1957). At the time, however, Bartlett regarded this as "rather an artificial evasion of the difficulty" and Bayesian textbooks such as Press (1989, p35) still warn that it may be impossible to conduct a meaningful hypothesis test of a simple null hypothesis against a diffuse alternative. The remedy usually recommended is to use a realistic subjective prior distribution over the space of $H_1$ instead.

However, a number of Reference Bayesians (including Jeffreys 1961, Smith and Spiegelhalter 1980, Aitkin 1991, O'Hagan 1995, Berger 1995 and Berger
and Pericchi 1996) have attempted to tackle the problem by devising proper priors that deliver plausible values for \( B \). A recent airing of diverse views on this topic can be found in the Discussion to O'Hagan (1995).

Formally, the generalized RPO, to be denoted here by \( G_{gr} \), can be defined by the requirement that \textit{when \( n \) is large and the observations are such that the posterior odds ratio, \( G_{gr} \), attains its maximum possible value conditional on \( n \), its leading term should be equal to \( k\sqrt{n} \).} This implies that

\[
1 = \lim_{n \to \infty} \left\{ \frac{\max(G_{gr})}{k\sqrt{n}} \right\} = \left( \frac{\pi_0}{\pi_1} \right)_{gr} \lim_{n \to \infty} \left\{ \frac{\max(B)}{k\sqrt{n}} \right\},
\]

where \( \left( \frac{\pi_0}{\pi_1} \right)_{gr} \) is the generalized reference prior odds ratio.

The above equation can be used wherever \( H_0 \) is precise and the prior density of \( H_1 \) in the region of \( H_0 \) is locally flat. (In the limit where \( H_1 \) is improper, \( \pi_0 \) will be infinitesimal and \( \pi_1 \) will be almost one, but the equation can still be used.) For any given \( n \), let \( y_0 \) be the value of \( y \) for which \( \text{Lik}(\theta_0; y_0) \) attains its maximum value. Then for any chosen value of \( k \) the generalized reference prior odds ratio is given by

\[
\left( \frac{\pi_0}{\pi_1} \right)_{gr} = \lim_{n \to \infty} \left\{ \frac{k\sqrt{n}}{\int_{\theta \in H_1} \text{Lik}(\theta; y_0) f_1(\theta) d\theta} \right\} \frac{\text{Lik}(\theta_0; y_0)}{\text{Lik}(\theta_0; y_0)}
\]

and the generalized RPO itself is
The proposed specific RPO, $G_{r}$, is the special case of $G_{r}$ for which $k = \sqrt{2}$.

**Example 3.1: Normal mean, variance known.**

Let the $n$ incoming observations $y = (y_1, y_2, \ldots, y_n)$ each be $N(\mu, \sigma^2)$ with $\mu$ the test parameter and $\sigma^2$ known. The two hypotheses are $H_0: \mu = \mu_0$ and $H_1: \mu$ is distributed uniformly over $(\mu_0 - C, \mu_0 + C)$ where $C \gg \sigma$.

Provided the true value of $\mu$ lies in the interval $(\mu_0 - C + 5\sigma, \mu_0 + C - 5\sigma)$, it is nearly certain that the posterior odds, $G_r$, will be closely approximated by

$\left(\frac{\pi_0}{\pi_1}\right)_{G_r} = 2C \sqrt{n/(2\pi \sigma^2)} \exp\left\{-n(\bar{y} - \mu_0)^2/(2\sigma^2)\right\}$,

where $\bar{y}$ is the mean of the $n$ observations, $y$. The general reference criterion is that

$max_{\gamma} G_{r} = 2C \sqrt{n/(2\pi \sigma^2)}(\pi_0/\pi_1)_{G_r} = k\sqrt{n}$,

implying that

$\left(\frac{\pi_0}{\pi_1}\right)_{G_r} = k\sigma \sqrt{\pi}/(C\sqrt{2})$

for any $C$. So $\lim_{C \to \infty} \pi_1$ is almost one, $\lim_{C \to \infty} (C\pi_0) = k\sigma \sqrt{\pi/2}$ and

$G_{r} = k\sqrt{n} \exp\{-n(\bar{y} - \mu_0)^2/(2\sigma^2)\}$.  

An immediately obvious choice for $k$ is unity, and this would indeed have the advantage that $-2 \ln G$, would (for the normal mean) be precisely equal to the Bayes information criterion (or twice the Schwarz criterion, Schwarz 1978). It corresponds to the situation where, given a single observation located exactly at $\mu_0$, the experimenter would be indifferent between $H_0$ and $H_1$. The reference value proposed here, however, is $k = \sqrt{2}$. Some convenient results that follow from that choice will be encountered in Sections 4 and 5.

Example 3.2: Normal variance, mean unknown.

Let the $n$ observations $y_i$ be distributed $N(\mu, \sigma^2)$ with $\mu$ initially distributed uniformly over the real line. The precise null hypothesis, $H_0$, is that $\sigma^2 = \sigma_0^2$ and the composite alternative, $H_1$, that $\sigma^2 \neq \sigma_0^2$, the alternative prior having the distribution $k_{u,e,\omega} \sigma^{-2a}$ between $e$ (arbitrarily small) and $\omega$ (arbitrarily large) so that $k_{u,e,\omega} = \left[ \int_e^\omega \sigma^{-2a} d\sigma^2 \right]^{-1}$.

The statistics $\bar{y} = n^{-1} \sum_{j=1}^n y_j$ and $s^2 = (n-1)^{-1} \sum_{j=1}^n (y_j - \bar{y})^2$ are sufficient for $\mu$ and $\sigma^2$ respectively. Integrating over $\mu$, the likelihood of $\sigma^2$ indexed on $s^2$ only may be written

$$L(\sigma^2; s^2) = \sigma^{-(n-1)} \exp[-(n-1)s^2/(2\sigma^2)].$$

The expectation of this expression over the parameter space of $H_1$ is

$$\int_{H_1} L(\sigma^2; s^2)f_1(\sigma^2)d\sigma^2 \equiv k_{u,e,\omega} \Gamma[(n-3+2a)/2] 2/[(n-1)s^2]^{(n-3+2a)/2}.$$
This approximation can be made arbitrarily close—in the sense that the ratio of the two sides can be made to approach unity—by choosing $\varepsilon$ to be small enough and $\omega$ to be large enough. With this approximation in mind, the Bayes Factor in favour of $H_0$ over $H_1$ is

$$B_a = \frac{\sigma_0^{2(a-1)}[(n-1)s^2/(2\sigma_0^2)]^{(n-3+2a)/2} \exp[-(n-1)s^2/(2\sigma_0^2)]}{k_{a,\varepsilon,\omega} \Gamma[(n-3+2a)/2]}.$$ 

This expression, and hence also the posterior odds ratio, attains its maximum value over $s^2$ when $(n-1)s^2/\sigma_0^2 = n - 3 + 2a$, in which case

$$\max_{s^2} B_a = \frac{\sigma_0^{2(a-1)}[(n-3+2a)/2]^{(n-3+2a)/2} \exp[-(n-3+2a)/2]}{k_{a,\varepsilon,\omega} \Gamma[(n-3+2a)/2]}.$$ 

Equating $\max_{s^2} B_a \left( \frac{\pi_0}{\pi_1} \right)$ asymptotically to $k\sqrt{n}$ and applying Stirling's formula, the generalized reference prior odds ratio is given by

$$\left( \frac{\pi_0}{\pi_1} \right)_{gr} = \frac{k\sqrt{n}}{\max_{s^2} B_a} = \frac{2k\sqrt{\pi n} k_{a,\varepsilon,\omega}}{\sigma_0^{2(a-1)}}$$

and the corresponding generalised RPO is

$$G_{gr} = \lim_{\varepsilon \to 0, \omega \to -\infty} \left\{ B_a \left( \frac{\pi_0}{\pi_1} \right)_{gr} \right\}$$

$$= \frac{2k\sqrt{\pi n}[(n-1)s^2/(2\sigma_0^2)]^{(n-3+2a)/2} \exp[-(n-1)s^2/(2\sigma_0^2)]}{\Gamma[(n-3+2a)/2]}.$$
Example 3.3: Binomial parameter

Let the probability of success on a single trial be $P$. Consider the test of $H_0: P = P_0$ against $H_1: P \neq P_0$, the prior density over the alternative parameter space being $\pi_i f_i(P) = \pi_i P^{\alpha-1}(1 - P)^{\beta-1} \Gamma(\alpha + \beta) / \Gamma(\alpha) \Gamma(\beta)$. Suppose $r$ successes are observed in $n$ trials. The posterior odds ratio is then

$$G = \frac{\pi_0}{\pi_1} B = \frac{\pi_0}{\pi_1} \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} \frac{\Gamma(n + \alpha + \beta)}{\Gamma(r + \alpha) \Gamma(n - r + \beta)} P_0^r (1 - P_0)^{n-r}.$$

This attains its maximum value over $r$ when \( \frac{r + \alpha}{n + \alpha + \beta} = P_0 \). In the limit as $n \to \infty$, this implies $r/n \equiv P_0$ and, since the leading term in $G_{gr}$ must be $k \sqrt{n}$,

$$\frac{G_{gr}}{\sqrt{n}} \equiv k \left( \frac{\pi_0}{\pi_1} \right)_{gr} \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} \lim_{n \to \infty} \left\{ \frac{\Gamma(n + \alpha + \beta)}{\Gamma(r + \alpha) \Gamma(n - r + \beta)} \frac{r'(n - r)^{n-r}}{n^{(2n+1)/2}} \right\}.$$

Applying Stirling's formula, we have

$$\left( \frac{\pi_0}{\pi_1} \right)_{gr} = k \sqrt{2\pi} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} P_0^{\alpha-0.5} (1 - P_0)^{\beta-0.5} \frac{r'(n - r)^{n-r}}{n^{(2n+1)/2}}.$$

and the generalized RPO is

$$G_{gr} = \left( \frac{\pi_0}{\pi_1} \right)_{gr} B = k \sqrt{2\pi} \frac{\Gamma(n + \alpha + \beta)}{\Gamma(r + \alpha) \Gamma(n - r + \beta)} P_0^{r+\alpha-0.5} (1 - P_0)^{n-r+\beta-0.5}.$$
Three cases having some claim to be noninformative are of special interest:

1. $\alpha = \beta = 1$, the uniform prior;
2. $\alpha = \beta = 0.5$, the Jeffreys prior (Jeffreys 1961, p125) and
3. $\alpha = \beta = 0$, the improper prior proportional to $P^{-1}(1 - P)^{-1}$.

Comparing the formulae for $G_{gr}$ in the three cases, the RPO may be shown to be asymptotically stable against these variations in $\alpha$ and $\beta$, provided only that $P_0$ is neither zero nor one. It is least stable when either $r$ or $n - r$ is small. A numerical example of this kind is given in Part II.

The expression for $(\pi_0 / \pi_1)_{gr}$ is not stable against changes in $\alpha$ or $\beta$ and tends to zero as $\alpha, \beta \to 0$. When $\alpha = \beta = 0.5$, however, it takes a particularly simple form, $k \sqrt{2/\pi}$, which is uniquely invariant over $P_0$ (cf. Jeffreys 1961, p188).

**Remark 3.1:** This Section has indicated how to define a reference prior odds ratio when the null hypothesis is precise and the alternative prior is composite. It has not so far indicated how reference prior probabilities should be assigned in more general situations or, in particular, how to ensure that the Four Models Paradox described in Section 2 can be avoided.

To an extent, each case would have to be treated on its own merits, but some general principles are clear. Suppose for instance that the models $M_{11}$ and $M_{12}$ are precise, that models $M_{21}$ and $M_{22}$ are composite, and further that $M_{11}$ is nested within $M_{21}$ and $M_{12}$ within $M_{22}$. Then the reference rule given in this Section can be used to define prior odds for $M_{11}$ over $M_{21}$ and $M_{12}$ over
$M_{22}$, but there will in general be an ambiguity affecting the prior odds favouring $M_{11} \cup M_{21}$ over $M_{12} \cup M_{22}$. One possibility is to accord each of these nested pairs the prior probability 0.5, but another is to regard the two precise hypotheses ($M_{11}$ and $M_{21}$) as equiprobable \textit{a priori} and allow this decision to determine the prior odds favouring $M_{11} \cup M_{21}$ over $M_{12} \cup M_{22}$. If there is nothing in the problem obviously pointing towards one or the other of these two possibilities, it would probably be best to analyse both of them and see whether any marked difference emerges.

4. THE POSTERIOR BAYES FACTOR

Concerned at the sensitivity of the Bayes Factor to the specification of the alternative prior, Aitkin (1991) introduced a statistic that he called the Posterior Bayes Factor (or PBF) and denoted by $A$. It differed from the conventional Bayes Factor in that the means of the likelihoods were taken over the posterior rather than the prior distributions, resulting in the expression

$$A = \frac{\int_{\theta \in H_0} \text{Lik}(y; \theta) f_0(\theta) d\theta}{\int_{\theta \in H_0} \text{Lik}(y; \theta) f_0(\theta) d\theta} / \frac{\int_{\theta \in H_1} \text{Lik}(y; \theta) f_1(\theta) d\theta}{\int_{\theta \in H_1} \text{Lik}(y; \theta) f_1(\theta) d\theta}$$

$$= \frac{E_0[\text{Lik}(\theta_0; y)]}{E_0[\text{Lik}(\theta_0; y)]} / \frac{E_1[\text{Lik}(\theta_1; y)]}{E_1[\text{Lik}(\theta_1; y)]}.$$

This statistic did not meet with favour from orthodox Bayesians, mainly because it was not really a Bayes Factor, except in so far as it corresponded to one special and highly informative prior (Fearn 1991), and because it used the
same data twice to make a single inference in such a fashion as to be
temporally incoherent (Cuzick 1991, Cox 1991). There is, however, a close
relationship between $A$ and the Realization Factor, $R$.

**Theorem 4.1:** Given that

1. $H_0$ is of the simple or precise form $\theta = \theta_0$,
2. $f_1(\theta)$ is flat,
3. there is a univariate sufficient statistic, $\hat{y}$, for $\theta$ and
4. $L(\theta; \hat{y})$ is origin invariant,

then $A$ and $R$ are equal.

**Proof:** Since $H_0$ is precise,

$$A = \frac{\text{Lik}(\theta_0; \hat{y})}{\text{E}_i\{\text{Lik}^2(\theta; \hat{y})\}/\text{E}_i\{\text{Lik}(\theta; \hat{y})\}},$$

and

$$R = \frac{\text{Lik}(\theta_0; \hat{y})}{\text{E}_i\{\text{Lik}(\theta; \hat{y})\}/\text{E}_i\{\text{Lik}(\theta; \hat{y})\}} \int \frac{\text{Lik}(\theta_0; \hat{y})}{\text{Lik}(\theta_0; \hat{y})} f_0(\hat{y}) d\hat{y}.$$

Since $L(\theta; \hat{y})$ is origin invariant and $f_1(\theta)$ is flat, $E_i\{L(\theta; \hat{y})\}$ is invariant
with $\hat{y}$ and can be cancelled out, leaving

$$R = \frac{\text{Lik}(\theta_0; \hat{y})}{\int \text{Lik}(\theta_0; \hat{y}) f_0(\hat{y}) d\hat{y}}$$

$$= \frac{\text{Lik}(\theta_0; \hat{y})}{\int \text{Lik}(\theta_0; \hat{y}) \int \text{Lik}^2(\theta_0; \hat{y}) d\hat{y}}$$

$$= \frac{\text{Lik}(\theta_0; \hat{y})}{\int \text{Lik}(\theta_0; \hat{y}) d\hat{y}/\int \text{Lik}(\theta_0; \hat{y}) d\hat{y}}$$
= \text{Lik}(\theta_0; \tilde{y}) \left[ \frac{\int_{\theta \in \mathcal{H}_0} \text{Lik}^2(\theta; \tilde{y}) f_1(\theta) d\theta}{\int_{\theta \in \mathcal{H}_0} \text{Lik}(\theta; \tilde{y}) f_1(\theta) d\theta} \right] \\
= A . \quad \Box

\textbf{Remark 4.1:} Since Conditions 3 and 4 of the Theorem are satisfied by the normal distribution, the asymptotic equality of $R$ and $A$ can be invoked whenever $H_0$ is precise, $f_1(\theta)$ is locally uniform and the Central Limit Theorem is operative. \quad \Box

Whenever Theorem 1 holds, the PBF can also be interpreted as the posterior odds corresponding to the situation where the prior distribution is identical with the reference prior defined above, except that the (infinitesimal) amount of prior probability associated with $H_0$ is chosen to ensure that $G_{gr} = A$. Such a prior may be viewed as reflecting the beliefs of a person who, faced with an experiment of size $n$ in which the posterior odds take their maximum value given $n$, regards those posterior odds as being $\sqrt{2}:1$ on. So $G_{gr} = A$ implies that $k$ has been chosen to be $\sqrt{2/n}$. If, however, $k$ has already been chosen to take the specific reference value, $\sqrt{2}$, (i.e. $G_{gr} = G_r = A\sqrt{n}$) then $G_{gr} = A$ only for the special case where $n = 1$. It might perhaps be regarded as unduly arbitrary to adopt a prior in which the experiment size, $n$, figures so prominently, but such a judgment could well be regarded as implicit in the (equally arbitrary) choice of the experiment size itself. Alternatively, one might argue that the alternative prior can only be neutral for one particular value of $n$, and that the natural value to choose is the size of the experiment.
There is, however, a substantial difference between the interpretation of a hypothesis test based on the use of the statistic $G$, (with $k = \sqrt{2}$) and that of one based on the use of $A$ (or, almost equivalently, of $R$). If $k = \sqrt{2}$ is used, there is a quite definite a priori degree of belief associated with $H_0$. While the value of the prior probability, $\pi_0$, associated with $H_0$ may itself be infinitesimal, the ratio of $\pi_0$ to the amount of prior probability associated with the region of the alternative hypothesis within, say, two observational standard deviations of the null hypothesis parameter value, $\theta_0$, is finite and of the order of unity. For the normal mean, $\mu$ (with variance, $\sigma^2$, known) the reference value of $\pi_0$ is equal to the infinitesimal amount of alternative prior probability contained in the interval $(\mu_0 - \sqrt{\pi}\sigma, \mu_0 + \sqrt{\pi}\sigma)$ or, equivalently, in an interval of length $2\sqrt{\pi}\sigma$ anywhere on the real line (Part II, Subsection 2.1).

If $A$ or $R$ is used, however, the implied degree of belief in $H_0$ is proportional to $1/\sqrt{n}$. This can only be compatible with Bayesian inference if $n$ is held fixed throughout the inference process. It would not be permissible, for instance, to stop the experiment halfway through and make a provisional inference based on the value of $A$ obtained from that half of the experiment. The two inferences taken together would be incoherent.

It therefore seems reasonable to say that the use of $A$ in the fashion recommended by Aitkin is not fully Bayesian in spirit. Instead of there being a fixed prior degree of belief in $H_0$, there is the implicit attitude that $H_0$ can never really be trusted, and that the only question at issue is whether or not there is already an experiment size large enough to demonstrate that it is false.
The similarity between this approach and that of the Frequentist statistician is obvious, and it is not surprising to see that Aitkin himself drew attention to the repeated sampling properties of $A$ (Aitkin 1991, Section 3).

5. THE CLASSICAL OR FREQUENTIST TEST STATISTIC

Under the conditions required for Theorem 1, $A$ and $R$ are interchangeable and, when considering relationships with the Frequentist test statistic, it is more convenient to think in terms of $R$. For the three examples in Section 3 it is straightforward to show that the choice of a uniform prior distribution over $H_1$ leads to a formula for $R$ which is proportional either to the density or to the actual distribution of the corresponding Classical test statistic under $H_0$.

For Example 1 (the normal mean) $R$ (i.e. $G_j/\sqrt{n}$) is proportional to the ordinate of $N(\mu_0, \sigma^2/n)$, the normal density of the Classical test statistic.

For Example 2 (the normal variance) the statistic $R$ or $G_j/\sqrt{n}$ may be interpreted as the ordinate of $\chi^2_{n-1+2a}$ standardized so as to have the expectation unity over repeated sampling when $\sigma^2 = \sigma^2_0$. Choosing $a = 0$, $R$ is the similarly standardized ordinate of the density of $\chi^2_{n-1}$. The Classical test statistic is also $\chi^2_{n-1}$, but with its ordinates now standardized so as to ensure that the value of its definite integral over the alternative parameter space is equal to one.
For Example 3 (the binomial parameter) $G$, $B$ and $R$ are all inversely proportional to $r!(n-r)!$ when $\alpha = \beta = 1$, i.e. when the alternative prior is uniform; and this is also the condition for proportionality to the distribution of the Classical test statistic.

While this correspondence may not hold in general, we may still assert that wherever the Central Limit Theorem can be invoked, the modified RB test (based on the use of $R$ as a test statistic and of a flat alternative prior) has a one to one correspondence with the Classical hypothesis test; although the latter focuses on the repeated sampling distribution of the observations and uses tail areas where the former focuses on the possible parameter values and uses ordinates. In discrete cases such as the binomial, however, where the meaningful tail areas are limited in number, this one to one correspondence is restricted to a finite number of possible comparisons.

There is, moreover, a further difference between the Classical test and the unmodified RB test based on $G$ (i.e. the one with $k = \sqrt{2}$). This is the factor $\sqrt{n}$, which reflects the rate at which $B$ tends to increase with the size of the experiment when $H_0$ is true. It corresponds to the progressive elimination of alternatives that are nearly but not quite equivalent to $H_0$, and to the consequent reduction in the amount of alternative prior probability able to support $H_i$ in any effective sense. It is also this factor that, combined with the hypersensitivity of $B$ to the choice of alternative prior, explains the well-known Lindley Paradox (Lindley 1957), namely that a Classical significance test for $H_0$ could be indicating $y$ to be significant at the 5% level, while at the
same time the posterior probability of $H_0$, given $y$, could be as high as 95%, even though the corresponding prior probability was quite small.

6. THE CHOICE OF INFORMATION CRITERION

Aitkin (1991, Section 3) showed that for a comparison between two nested hypotheses with $v$ additional parameters in $H_1$, the multivariate version of $A$ implied a particular form of information criterion within a general class of penalized likelihood ratio test statistics discussed by Smith and Spiegelhalter (1980). Using $\lambda$ to denote the usual likelihood test statistic, this general class may be written $A(m) = \lambda - mv$ and, since $-2 \ln A = \lambda - v \ln 2$, the choice of multivariate $A$ as a test statistic implies the choice $m = \ln 2 = 0.693$. This is a smaller value for $m$ than is customarily used in practice. For comparison, the Akaike Information Criterion or AIC uses $m = 2$ and the Bayes Information Criterion [defined as twice the Schwarz Criterion, $S$, (Schwarz 1978)] uses $m = \ln n$, which is greater than 2 when $n \geq 8$.

The choice $m = \ln 2$ results in decisions to favour complex models to a counterintuitive extent. To see this, suppose that there were $v$ potential regressors for a given regressand, that (unknown to the experimenter) they were all spurious, and that each in turn was to be judged as either "real" or "spurious" depending on whether or not the value of $\lambda$ obtained by adding each candidate in turn to the equation exceeded $\ln 2$ (or 0.693). Since the probability that $\chi^2$ exceeds 0.693 is about 0.43, an expected 43 per cent of these spurious parameters would be judged to be "real", regardless of the size of the experiment. If they were to be tested as a group, the likely outcome
12 In 2 is just greater than 3/4, so the entire group would be judged “real” with a probability of more than 75 per cent.

By contrast, the use of $G_r$ in place of $A$ leads to the choice $m = \ln(2n)$, which is slightly more parsimonious than the BIC itself, exceeding the value 2 when $n \geq 4$. Moreover, Kass and Wasserman (1995) have recently suggested that for the BIC the relevant value of $n$ in the formula for $m$ is not the number of vector observations but the number of vector components in the observations, $nv$. If this is correct, the use of $G_r$ should lead to the choice $m = \ln(2nv)$, which is more parsimonious still. However the notion that this result should hold even when each component within the matrix of observations, $Y$, is perfectly correlated with every other is counterintuitive, and it seems wise to suspend judgment on this issue for the time being.

7. THE BAYESIAN–FREQUENTIST INTERFACE

The author of this thesis was trained in the Classical tradition and, although he has long been impressed by the Reference Bayesian approach, he has almost equally long been concerned at its apparent inability to provide an objective and universally acceptable test of a precise null hypothesis against an ideally improper alternative. One consequence has been that the gulf between the Bayesian and the Classical approaches to hypothesis testing has appeared virtually impassable for four decades or more, with only Aitkin (1991) giving any hint that the two might eventually be capable of reconciliation. This is unfortunate, as the two appear to have complementary advantages.
The objectivity of the Frequentist approach appeals to the scientist, the lawyer, and indeed to everyone who has to present inferences from data in a neutral fashion. Its weakness lies in the seeming irrelevance of its inferences to everyday life. Alexander (1994, p25) makes the pertinent admission that "... what people really would like to know is the probability that the population value is in the confidence interval, which isn't exactly what we non-Bayesians can tell them." I am myself concerned that non-statisticians so often find confidence intervals difficult to comprehend and also that, once they have understood them, they sometimes doubt whether they are useful.

The desired rapprochement would not and should not result in a "Bayes—non-Bayes compromise", as appears to have been suggested first by Jack Good (Barnard 1991, p128). The Bayesian approach does not admit of any compromise. But it does need a fully comprehensive description of how to proceed within the Reference Bayes approach. Topic B of this thesis provides a step in that direction: it shows that a precise null hypothesis can indeed be tested against a composite alternative in a neutral fashion, even if that alternative is improper; also, that the procedures required to do this can be closely related to the corresponding Frequentist test statistics. If the objectivity enjoyed by the Frequentists can be fully incorporated into the Reference Bayes approach, and if this enables Frequentist results to be given an easily comprehensible Bayesian interpretation, both the general public and the statistical profession will benefit substantially.
REFERENCES


Lindley, D.V., (1976), Comment in the discussion of Stone (1976), 120-121.


