Topics in empirical Bayesian analysis

by

Robert Christian Foster

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Mark Kaiser, Major Professor
Petruţa Caragea
Daniel Nettleton
Jarad Niemi
Daniel Nordman

Iowa State University
Ames, Iowa
2016

Copyright © Robert Christian Foster, 2016. All rights reserved.

DEDICATION

I would like to dedicate this dissertation to my mother, without whose constant love and support I would not have been able to complete it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank all my friends that supported me throughout graduate school — Danielle and Jonathan, Dan F., Brian, Dan A. and Julie, Laura, Lisa, Marcela, Kristian and Chris, and Adam and Arielle. Without their support I likely would not have completed this dissertation.

I would also like to thank the department of statistics at Iowa State University for continuing to support me, though I was not always the ideal student.

I would like to thank my committee of Dr. Jarad Niemi, Dr. Daniel Nettleton, Dr. Daniel Nordman, and Dr. Petruţa Caragea for their patience in working with me.

Lastly, I would like to thank my Dr. Kaiser for all the illuminating talks that we have had in the process of creating this dissertation, and for the many helpful comments and suggestions he has given.

ABSTRACT

While very useful in the realm of decision theory, it is widely understood that when applied to interval estimation, empirical Bayesian estimation techniques produce intervals with an incorrect width due to the failure to incorporate uncertainty in the estimates of the prior parameters. Traditionally, interval widths have been seen as too short. Various methods have been proposed to address this, with most focusing on the normal model as an application and many attempting to recreate, either naturally or artificially, a hierarchical Bayesian solution. An alternative framework for analysis in the non-normal scenario is proposed and, for the beta-binomial model, it is shown that under this framework the full hierarchical method may produce interval widths that are shorter than empirical Bayesian interval widths. Furthermore, this paper will compare interval widths and frequentist coverage for different Bayesian and non-Bayesian interval correction methods and offer recommendations. This framework may also be extended to the larger natural exponential family with quadratic variance functions, of which the beta-binomial model is a member, and general properties of NEFQVF distributions are given, with a specific application of the gamma-Poisson model. A class of prior is introduced as a limiting state of the framework that, in the hierarchical setting where the shrinkage coefficient is known, extends the well-known conjugacy of NEFQVF families to the hierarchical setting in an approximate way, and intervals are constructed using a refined empirical Bayesian interval correction technique that produce an alternative comparison basis. Coverage and interval widths are shown for this technique for the beta-binomial and gamma-Poisson models. Both produce near-nominal coverage and compare favorably to the full hierarchical solution calculated using MCMC.

As a second topic, a new Bayesian and empirical Bayesian estimate of a baseball team's "true" winning percentage is introduced. Common methods for estimating this "true" winning percentage, such as the pythagorean expectation or pythagenpat system, rely on the total number of runs scored and allowed over a period of time. A new estimator is proposed that uses independent zero-inflated geometric distributions for runs scored and allowed per inning to determine a winning percentage. This estimator outperforms methods based on total runs scored and allowed in terms of mean-squared error using actual win totals. Interval estimation for this estimator is directly shown using frequentist or Bayesian techniques. Empirical Bayesian techniques are shown as an approximation to the full hierarchical solution. Selected interval widths are compared using all three methods, with slight differences in the shrinkage amount given a full season's worth of data.

CHAPTER 1. INTRODUCTION AND OUTLINE

1.1 Bayesian and Empirical Bayesian Methods

Bayesian and empirical Bayesian methods have found a wide use within statistical practice. Modern usage of empirical Bayesian methodologies has ranged from small-area estimation problems to large sample analysis of microarray data, while traditional Bayesian methods have expanded into a rich practice in hierarchical modeling. This dissertation presents new research within the field of empirical Bayesian estimation and its connection to Bayesian estimation and frequentist estimation. Three separate applications of Bayesian and empirical Bayesian methodologies are introduced, focusing on practice, theory, and application.

1.2 Dissertation Organization

In the first chapter, focusing on practice, a brief literature review discusses various methods of correcting the widths of empirical Bayesian intervals, with the implied reasoning of the correction method as to the correct basis of comparison indicated. A new framework is introduced for the beta-binomial model that provides an intermediate stage between empirical Bayesian and hierarchical Bayesian methods which matches expectations of prior distributions to estimates of prior distributions, and some properties are shown that connect this framework to some of the given correction techniques. A specific application using medical data is shown in which empirical Bayesian interval widths exceed those of a “noninformative” hierarchical Bayesian analysis, and it is further shown that in this framework, it is entirely reasonable that an empirical Bayesian interval width may exceed a hierarchical Bayesian interval width. Simulations are conducted to determine what types of data sets are likely to have this phenomenon present itself, and to compare the properties of the intermediate priors and full hierarchical priors to the non-hierarchical correction methods for both known and unknown shrinkage amounts, with generally positive results.

In the second chapter, focusing on theory, the ideas from the first chapter are extended to the natural exponential family with quadratic variance function (whose properties are briefly reviewed), and a general form of the framework from the first chapter is introduced in the case with known shrinkage amount. Taking the limiting case of this framework leads to a class of hyperpriors which, in the case of known shrinkage amount, produce a form of strong approximate conjugacy for hierarchical models, and which may be used as a basis for comparison of empirical Bayesian interval widths in some scenarios. A specific case of the gamma-Poisson model is shown, and simulated data shows empirical Bayesian interval widths that exceed hierarchical Bayesian interval widths. Finally, the strong approximate conjugacy property is used to refine a method for correction of empirical Bayesian intervals widths described in Morris (1988), and simulations show the refined method performs well in terms of coverage and interval length, with results close to those of the full hierarchical model.

Lastly, an application of Bayesian and empirical Bayesian methodologies is given. A class of winning percentage estimators for baseball teams is briefly reviewed, along with efforts of modeling run scoring distributions, and a new estimator is introduced that relies on parametric modeling of run scored and allowed distributions with a zero-inflated geometric distributions. This winning estimator compares well to the existing estimators when fit using maximum likelihood techniques in terms of root mean squared difference from the actual win totals, and interval estimation (generally not attempted for winning percentage estimators) is shown using maximum likelihood and Bayesian estimation. Empirical Bayesian estimation is shown as an alternative to the full hierarchical Bayesian model, and offers a computationally simpler analysis.

CHAPTER 2. EMPIRICAL BAYESIAN INTERVAL WIDTHS

2.1 Introduction and Literature Review

The idea of “empirical Bayes” is not a precisely defined concept within modern statistics. Some consider it a specific technique, others consider it a class of techniques, and yet others consider it to be a philosophical approach to data analysis. Generally the idea revolves around empirical bayes as a sort of “pseudo-Bayes” — a way to, as Carlin and Louis (2000) state, “compromise between the frequentist and Bayesian methods.”

Though more commonly viewed through a decision-theoretic framework, intervals may be based on empirical Bayesian estimates. Conventional wisdom states that these intervals are necessarily too narrow due to underestimation of the variance; however, these comparisons are often based upon idealized scenarios, and usually only the Gaussian-Gaussian model.

2.1.1 Modern Empirical Bayes

The dual work of Herbert Robbins and Carl Morris & Brad Efron has led to the development of two branches of empirical Bayes theory. The first, nonparametric empirical Bayess, follows in the Robbins tradition of assuming that a common prior $G(.)$ exists, but the form is unknown, and so inference proceeds based off of the marginal density. The second branch, known as parametric empirical Bayes, builds off of Morris and Efron’s 1975 work connecting Stein’s estimator to Bayesian procedures. Parametric empirical Bayes will be the focus of this article.

The parametric empirical Bayes model will be as follows: suppose there are observations $y_1, ..., y_k$ (note that y_i may be a vector or a sufficient statistic) from some parametric density $f(y_i|\theta_i)$. A formal Bayesian analysis would choose some common prior $G(\theta_i)$, $i = 1, 2, ..., k$, and use Bayes' rule to derive the posterior distributions and an estimates for the θ_i. In parametric empirical Bayes, however, a prior distribution $G(\theta_i|\eta)$ indexed by some parameter η is used (and note that η may be a vector). An estimate $\hat{\eta}$ is provided by the marginal density $m_G(y_i|\eta) = \int f(y_i|\theta_i)G(\theta_i|\eta)d\theta_i$, usually through a common estimation procedure such as the method of moments or the method of maximum likelihood, and the posterior distributions for θ_i are derived using $G(\theta_i|\hat{\eta})$ as a common prior.

The problem is formulated as

$$\begin{aligned} y_1, y_2, ..., y_k &\overset{indep}{\sim} f(y_i|\theta_i) \\ \theta_i &\overset{iid}{\sim} G(\theta_i|\hat{\eta}) \end{aligned} \tag{2.1}$$

When working in an ideal empirical Bayesian setting (such as the Gaussian/Gaussian model), a 95% empirical Bayesian confidence interval might be given by

$$E[\theta_i|y_i, \hat{\eta}] \pm 1.96\sqrt{Var(\theta_i|y_i, \hat{\eta})} \tag{2.2}$$

It is commonly understood, however, that this is "naive" in the sense that it does not incorporate uncertainty in the estimated parameter or parameters $\hat{\eta}$ of the prior distribution. Under the parametric empirical bayes model, a standard two-stage variance decomposition of $Var(\theta_i|\boldsymbol{y})$ is given by

$$Var(\theta_i|y_i) = E_{\eta|y_i}[Var(\theta_i|y_i, \eta)] + Var_{\eta|y_i}[E(\theta_i|y_i, \eta)] \tag{2.3}$$

Carlin and Louis (2000) note that the quantity $\sqrt{Var(\theta_i|y_i, \hat{\eta})}$ in equation (2.2) approximates the first term — it does not, however, address the second term in equation (2.3), which Carlin and Louis (2000) identify as "the posterior uncertainty about η."

In the non-Gaussian setting, where the posterior distribution is not symmetric, intervals for the parameters θ_i may be produced through quantiles of the posterior distribution for θ_i — and though not taking the same form, share the same issue of not accounting for the posterior uncertainty regarding $\hat{\eta}$.

2.1.2 Corrected Intervals

In order to accurately assess the confidence level, a precise notion of "confidence" for empirical Bayes intervals must first be constructed. Carlin and Gelfand (1991) define $t_\alpha(\mathbf{y})$ as a $(1-\alpha) \times 100\%$ confidence interval for θ_i if

$$P_\eta(\theta_i \in t_\alpha(\mathbf{y})) \approx 1-\alpha \tag{2.4}$$

Morris (1983a) has a similar definition, but with $\geq 1-\alpha$. Carlin and Louis (2000) refer to this type of interval as an unconditional confidence interval — that is, it requires confidence over the variation in both the θ_i and the data $\mathbf{y}$.

Carlin and Gelfand (1991) note that some consider this statement weak, and suggest that "a probability statement which offers conditional calibration given an appropriate data summary" would likely be preferred. They define an alternative interval as $t_\alpha(\mathbf{y})$ to be a $(1-\alpha) \times 100\%$ confidence interval for θ_i if

$$P(\theta_i \in t_\alpha(\mathbf{y}) | b(\mathbf{y}) = b) \approx 1-\alpha \tag{2.5}$$

where $b(\mathbf{y})$ is an appropriate summary statistic of $\mathbf{y}$ (even up to taking $b(\mathbf{y}) = y_i$). This is known as a conditional empirical bayes confidence interval, since it requires confidence strictly over the variation in the θ_i. Essentially, demanding conditional coverage requires that given an appropriate summary statistic, any set of empirical Bayesian intervals calculated from that statistic has nominal coverage, while demanding unconditional coverage allows the conditional coverage given each summary statistic to vary so long as the coverage over all possible summary statistics is nominal. This paper will focus on estimating unconditional coverage. Multiple procedures have been proposed to adjust the coverage so as to reach nominal coverage rates.

2.1.3 Bias-Corrected Intervals

One approach to producing "correct" intervals — and one of the few not mimicking a hyperprior solution — is found in Efron's comments on Laird and Louis (1987) and further explained in Carlin and Gelfand (1990), as a method to conditionally correct the bias from the empirical Bayes procedure.

Taking the general empirical Bayes setup as described in equations (2.1) but first supposing that there exists a true value of η which is known, empirical Bayes confidence intervals are calculated by first taking quantiles from the posterior density of the θ_i:

$$(q_{\alpha/2}(y_i, \eta), q_{1-\alpha/2}(y_i, \eta))$$

where $q_\alpha(y_i, \eta)$ is defined by

$$P(\theta_i \leq q_\alpha(y_i, \eta) | \theta_i \sim p(\theta_i | y_i, \eta)) = \alpha$$

That is, they are the posterior quantiles from the posterior distribution that uses the true value of η. Of course, in practice the true value of η is unknown and must be estimated with $\hat{\eta}$ — it is then possible to define

$$r(\hat{\eta}, \eta, y_i, \alpha) = P(\theta_i \leq q_\alpha(y_i, \hat{\eta}) | \theta_i \sim p(\theta_i | y_i, \eta))$$

That is, the probability that θ_i (on the true posterior using η) is less than sample quantile taken from the estimated posterior using $\hat{\eta}$. Since $\hat{\eta}$ is random, the expected probability can be taken

$$R(\eta, y_i, \alpha) = E_{\hat{\eta}|y_i,\eta}[r(\hat{\eta}, \eta, y_i, \alpha)] \tag{2.6}$$

This is the average probability (over $\hat{\eta}$) that θ_i is less than the sample quantile taken from the estimated posterior using $\hat{\eta}$ with lower probability α. The expectation in equation (2.6) does not actually have to be close to the nominal α. However, it is possible to solve

$$R(\eta, y_i, \alpha') = \alpha$$

for α' — that is, there exists a nominal value α' that, when used to take sample quantiles, produces expected lower probability α on the true posterior. Using α' when taking quantiles from the empirical Bayesian posterior, then, could correct for the underestimation of the variance.

In practice, however, η is not know, so instead the estimate $\hat{\eta}$ is treated as the true value, and the quantity

$$R(\hat{\eta}, y_i, \alpha') = \alpha$$

is solved for α'. Again, using α' in place of α when taking quantiles from the posterior distribution for θ_i can be used to correct for the underestimation of the variance.

In the idealized normal-normal scenario, this correction can be calculated analytically. In other situations, it may have to be solved numerically, or a bootstrap method may be used to solve the unknown quantities. Theoretical details may be found in Carlin and Gelfand (1990) and practical implementation through a bootstrap approach may be found in Carlin and Gelfand (1991).

2.1.4 Morris Intervals

Most correction approaches attempt to place or mimic the results of a full hyperprior $h(\eta)$ on η. In Morris (1983b), Carl Morris uses a model with $y_i \sim N(\theta_i, \sigma^2)$ (σ^2 known) and $\theta_i \sim N(\mu, A)$ with a flat improper prior on μ and a $Unif[0, \infty)$ prior on A to derive intervals based on the full hierarchical Bayesian procedure. These intervals are centered around the empirical Bayesian estimate

$$\hat{\theta}_i = (1 - \hat{B})y_i + \hat{B}\bar{y}$$

Calculations for the posterior variance are complicated and do not extend well to the case with non-constant variance σ_i^2, so Morris (1983c) proposes an approximation by

$$\hat{\theta}_i \pm z_{\alpha/2}\sqrt{\sigma^2\left(1 - \frac{k-1}{k}\hat{B}\right) + \frac{2}{k-3}\hat{B}^2(y_i - \bar{y})^2}$$

where

$$\hat{B} = \frac{(k-3)\sigma^2}{\sum(y_i - \bar{y})^2}$$

Based on simulations in Carlin and Louis (2000), Morris's method produces empirical Bayes intervals that achieve or exceed the nominal 95% level in the normal situation, with a width that is shorter than the hierarchical Bayesian method. However, this formula is based explicitly on the assumption of a normal-normal model. Similar intervals are discussed in Efron (2010).

Morris extended the idea, if not the same equational form, of hierarchical Bayesian calculations in Morris (1988) to apply to empirical Bayes estimators when working with the natural exponential family with quadratic variance function, using a naive, moment-based estimator for η and using edgeworth expansions to account for sources of uncertainty. In the case of the normal distribution, this method worked very well and holds the standard 95% confidence property. For non-normal data, or for confidence intervals derived instead from quantiles of the posterior distribution of empirical Bayes, it is uncertain if the properties also hold, and Morris suggested that such can only be checked by Monte-Carlo techniques. In addition, Morris's technique assumes $n_i = n$ for all i, and this is admittedly uncommon in most applications. This would complicate his analysis in the normal case, and moreso in the non-normal case.

2.1.5 Bootstrap Approach

More commonly, the prior is estimated by bootstrap methodology. Laird and Louis (1987) discuss three types of bootstrap samples, with classification depending on assumed knowledge of the forms of $f(y_i|\theta_i)$ and $G(\theta|\hat{\eta})$. For the third type of bootstrap — which fits the parametric empirical Bayesian framework and will be used in this paper — a sample θ^* is produced from $G(\theta|\hat{\eta})$, which is subsequently used to produce a sample from the distribution $f(y_i|\theta_i^*)$. An estimate η^* is then calculated using the same techniques used on the marginal distribution $m_G(y_i|\eta)$ as before. For any of the bootstrap sampling techniques, intervals can then be calculated by mimicking the hyperprior distribution calculation — in particular, for N bootstrap samples, define

$$f_G^*(\theta_i|y_i, \hat{\eta}) = \frac{1}{N} \sum_{j=1}^{N} f_G(\theta_i|y_i, \eta_j^*)$$

where η^* is the estimated value of η for each bootstrap sample and $f_G(\theta_i|y_i,\eta)$ is the posterior distribution of θ_i given data y_i and prior parameters η. The distribution $f_G^*(\theta_i|y_i,\hat{\eta})$ represents an artificially created posterior distribution for θ_i (centered around the same value as the non-bootstrapped estimate) incorporating the uncertainty in $\hat{\eta}$ estimated by the bootstrap method. Equal-tailed confidence intervals for θ_i can then be found by solving

$$\frac{\alpha}{2} = \int_{-\infty}^{C_L} f_G^*(\theta|y_i,\hat{\eta}) = \int_{C_U}^{\infty} f_G^*(\theta|y_i,\hat{\eta})$$

This often must be calculated numerically, although in the normal-normal case the third type of bootstrap interval is equivalent to a flat hyperprior on η when the prior variance is known. In the normal-normal case with unknown prior variance, the bootstrap posterior can not match a hyperprior Bayesian posterior. Simulation results show that the third type of bootstrap intervals compare very well to the naive bootstrap intervals.

2.1.6 Hyperprior Approach

Another approach is proposed by Deely and Lindley (1991). Known as "Bayes empirical Bayes", a clever sequence of substitutions and applications of Bayes rule is used to determine that for a given hyperprior $h(\eta)$, the posterior distribution of θ_i independent of η may be calculated as

$$p(\theta_i|\mathbf{y}) = \frac{\int p(\theta_i|y_i,\eta)\prod_{i=1}^{k} m_G(y_i|\eta)h(\eta)d\eta}{\int \prod_{i=1}^{k} m_G(y_i|\eta)h(\eta)d\eta}$$

Though equation (2.1.6) is usually not numerically tractable, derivations of forms for the exponential family are given by Deeley and Lindley, and an example is given using an exponential-Poisson model with an improper hyperprior on η. Quantiles could be taken from this density to form an interval. Walter and Hamedani (1987) give a form of Bayes empirical Bayes estimation for a binomial probability using orthogonal polynomials and Walter and Hamedani (1991) extend this approach to the natural exponential family with quadratic variance functions.

One key difference in the theoretic formulation between the Morris and bootstrap methods and the Bayes empirical Bayes method is that the Morris and bootstrap methods essentially treat the empirical Bayesian estimates as correct and attempt to adjust the posterior distribution around them to achieve nominal coverage. No claim is made, however, that the posterior means of the Bayes empirical Bayes densities matches the empirical Bayes posterior means. This subtle difference belies a fundamentally different approach to the connection between empirical and full Bayesian methodologies.

2.2 The Empirical Bayes Problem for the Beta-Binomial Model

2.2.1 The Beta-Binomial Model

Historically, the development of parametric empirical Bayesian methods, including correction methods discussed in Section 2.1, has focused upon the normal-normal model, and results derived apply well to that scenario. This is particularly true in the area of interval estimation, where, apart from positive-part corrections, likelihood estimates are functions of basic moments, and prior distributions that yield hierarchical Bayesian estimates which are equivalent to the empirical Bayesian estimates exist and can be determined.

Less effort has been expended on the application of the previous techniques to models other than the normal-normal; or, if they are mentioned, it is only as a theoretical application of the technique that is not studied — Carlin and Gelfand (1991) gives technical details for parametric bootstrap estimation of the bias correction method applied to the beta-binomial model without providing a specific example or study to illustrate the method.

Empirical Bayesian methodologies are applicable in this situation, and in fact, the original example showing the effectiveness of empirical Bayesian methods in Efron and Morris (1975) was predicting baseball batting averages after transformation so that the normal-normal model could be applied. For the normal-normal model, the idea that a simple widening of the empirical Bayesian corrects the empirical Bayesian problem of uncertainty in the hyperprior parameters is taken as a given. Whether this issue can also be taken for granted in a non-normal scenario, and whether correction methods are necessary, has generally not been investigated.

2.2.2 Beta-Binomial Model Formulation and Estimation Form

The beta-binomial model is now considered:

$$\begin{aligned} y_i|\theta_i &\overset{indep}{\sim} Bin(\theta_i, n_i) \\ \theta_i &\overset{iid}{\sim} Beta(\mu, M) \end{aligned} \tag{2.7}$$

for $i = 1, 2, ..., k$. In place of the traditional α, β parametrization, the transformation $\mu = \alpha/(\alpha + \beta)$ and $M = \alpha + \beta$ will be applied, representing the mean of the beta distribution and a parameter that controls (but is not equal to) the variance, respectively — M will be referred to as a “variance parameter.” All beta distributions will be defined in terms of μ and M unless otherwise noted. This parametrization of the beta distribution then has mean and variance

$$\begin{aligned} E[\theta_i] &= \mu \\ Var(\theta_i) &= \frac{\mu(1-\mu)}{M+1} \end{aligned}$$

An alternative parametrization that will be used when M is unknown is

$$\theta_i \sim Beta'(\mu, \phi) \tag{2.8}$$

where μ is as before and $\phi = 1/(\alpha + \beta + 1) = 1/(M + 1)$ is the dispersion parameter of the beta distribution. In this paper, the $Beta'(\mu, \phi)$ distribution with ϕ is used only for modeling purposes and no priors are defined using this parametrization.

A simple Bayesian analysis would choose values for μ and M (for example, $\mu = 0.5$ and $M = 1$ yields the Jeffrey’s prior) and proceed using the well-known conjugacy of the beta prior for the binomial distribution. The form of the Bayesian estimator given by the expectation of the resulting posterior is then

$$\hat{\theta} = \mu + \left(1 - \frac{M}{M+n_i}\right)\left(\frac{y_i}{n_i} - \mu\right) \tag{2.9}$$

The empirical Bayesian method would be to estimate the parameters μ and M by some method, such as the method of moments or marginal maximum likelihood, and use equation (2.9) above with $\hat{\mu}$ and $\hat{M}$ in place of μ and M. The empirical Bayesian estimator is then

$$\hat{\theta} = \hat{\mu} + \left(1 - \frac{\hat{M}}{\hat{M} + n_i}\right)\left(\frac{y_i}{n_i} - \hat{\mu}\right) \tag{2.10}$$

Intervals may be constructed by drawing quantiles from the posterior distribution for θ_i

$$(q_{\alpha/2}(y_i, \hat{\mu}, \hat{M}), q_{1-\alpha/2}(y_i, \hat{\mu}, \hat{M}))$$

Another option is to perform a full hierarchical Bayesian analysis by choosing hyperpriors for the parameters μ and M (or equivalently ϕ) and conducting the full analysis through Markov-Chain Monte Carlo techniques, though technical details of the MCMC procedure will not be covered here. The resulting $\hat{\theta}$ estimates can not, in general, be calculated as solutions of closed-form equations.

2.2.3 Framework for Comparison

The traditional folklore of empirical Bayesian analysis is that it underestimates the variance by using the data twice, and underestimation of the variance takes the form of the standard two-stage variance decomposition in equation (2.3). In addressing this, there generally has been only one methodology — that of Morris (1983b) which considers the appropriate correction to be that obtained through a fully Bayesian analysis with a prior distribution such that

$$E[\theta_i|y_i] \approx \hat{\theta}_{iEB}$$

where $\hat{\theta}_{iEB}$ is given by standard empirical Bayesian estimators, as in equation (2.2.2). Here, the posterior expectation of the fully Bayesian procedure yields the same expected value as the empirical Bayesian estimators — the exception being the "Bayes empirical Bayes" analysis of Deely and Lindley (1991), which gives an example of computing posterior expectations in exponential families with very specific priors and hyperpriors.

2.2.4 Additional Framework

The Morris-type framework is particularly appealing in that it does not require the comparison of differences in posterior estimates caused by differences in priors; rather, it considers the empirical Bayes estimates as "correct" and defines the true interval width to be the interval width resulting from a full Bayesian analysis that produces those estimates. Furthermore, there is no need to compare different estimation techniques as there is, essentially, only one estimator.

This paper will propose and investigate additional frameworks more similar to the framework of Deeley and Lindely — first, rather than matching empirical Bayesian estimates of θ_i to posterior Bayesian estimates of the θ_i, one may match estimates $\hat{\mu}$ and/or $\hat{M}$ to the means of hyperprior distributions in a hierarchical Bayesian analysis. In this way, the full prior on the θ_i given by integrating out over the priors at each stage will have expected value matching the expected value of the empirical Bayesian prior distributions — and so the interval widths may be compared to a full Bayesian analysis with a sufficiently diffused prior rather than posterior. This has the added benefit of using proper prior distributions, although this method, similarly to the Bayes empirical Bayes method of Deely and Lindley (1991), will not necessarily lead to posterior θ_i estimates that are equivalent to empirical Bayesian estimates.

In the general hierarchical Bayesian setting, this will be proposed as

$$\begin{aligned} y_i &\overset{indep}{\sim} f(y_i|\theta_i) \\ \theta_i &\overset{iid}{\sim} G(\theta_i|\eta) \\ \eta &\sim h(\eta|\hat{\eta}) \end{aligned} \tag{2.11}$$

where $\hat{\eta}$ is the estimate used in the empirical Bayesian analysis. As an example, in the normal-normal scenario with known prior variance τ^2, this would be proposed as

$$\begin{aligned} y_i &\overset{indep}{\sim} N(\theta_i, \sigma^2) \\ \theta_i &\overset{iid}{\sim} N(\mu, \tau^2) \\ \mu &\sim N(\hat{\mu}, \tau_0^2) \end{aligned}$$

The empirical Bayesian estimator may then be constructed as the resulting posterior taking the hyperprior variance τ_0^2 to zero and creating a degenerate distribution at $\hat{\mu}$ with probability one, and the flat hyperprior $h(\mu) \propto 1$ that Morris (1983b) uses to produce interval corrections can be created by taking the un-normalized density as τ_0^2 approaches infinity.

Aside from producing an intermediate connection between empirical and hierarchical Bayesian estimates, the necessity of this is that in models other than the normal-normal, there may exist multiple methods to estimate the parameters of the prior distribution — and not all of these methods will have an extant prior that yields posterior expectations equal to empirical Bayes estimates without using a highly informative prior. One may be approximated using the bootstrap method of Laird and Louis (1987), but there's no reason to suggest that this is necessarily the correct thing to do or that it is the only possible choice for comparison — there may exist a hierarchical Bayesian prior yielding nominal frequentist coverage with shorter intervals.

As data for comparison, two different sources will be used:

1. Existing data that has been analyzed with empirical Bayesian techniques

2. Data randomly generated from a beta-binomial model

Coverage assessments are possible for the second data source, and interval widths may be contrasted between empirical Bayesian and hierarchical Bayesian methodologies for both data sources.

2.3 Analysis

2.3.1 Interval Widths

A key question that illustrates the inadequacy of the current comparison framework is whether it is possible under reasonable circumstances to have an empirical Bayesian interval that is, in fact, wider than the hierarchical Bayesian width and if so, by how much. This is an issue generally not considered under the idealized normal-normal model, as the corrected posterior retains the traditional bell shape, but with an increased posterior variance. However, it is an issue that can occur for the beta-binomial model under the proposed framework, and particularly occurs due to the presence of several consistent estimators for the parameters μ and M of the underlying beta distribution.

2.3.2 Estimation Procedures

The two procedures considered for estimation of the parameters in the beta-binomial model are the method of moments and the method of maximum likelihood. Both procedures were applied to the marginal density of the beta-binomial model, more commonly known as the beta-binomial distribution. For method of moments, both a “naive” method that only requires plugging in of values and a more sophisticated moments procedure that takes an iteratively weighted average were considered. Details of the estimation procedures may be found in Appendix A.

When all binomial sample sizes are equal (the case that will typically be addressed in this paper), the naive moments estimator produces the same estimates as the iterated moments estimator, and will simply be referred to as the method of moments estimate. When sample sizes differ, the iterated moments estimator produces estimates that tend to be closer in value to the estimates produced by maximum likelihood.

2.3.3 Practical Example

The method used for prior estimation can and will affect resulting interval widths - an example comes from Young-Xu and Chan (2008). In trials of antifungal medicine terbinafine, $N = 41$ treatments arms were considered. The response y_i in arm i is the number of patients (out of n_i total) that withdrew from the trial due to adverse reactions to the medication. These data are presented in Table B.1 in Appendix B.

Empirical Bayesian estimates and intervals were calculated using the naive method of moments, an iteratively weighted method of moments, the method of maximum likelihood, and a Bayesian procedure with priors

$$p(\mu) \propto \frac{1}{\mu(1-\mu)}$$

$$\phi \sim Beta(0.5, 1)$$

where ϕ is as in equation (2.8). The full set of Bayesian priors was chosen to represent a noninformative Bayesian solution.

Selected empirical Bayesian estimates using a naive unweighted method of moments (NMM), iterated method of moments (IMM), maximum likelihood (MLE), and full hierarchical Bayesian (HB) estimates, rounded to three decimal places, are given below in Table 2.1. The full table is given in Appendix B.

Table 2.1 Selected Empirical Bayesian and Hierarchical Bayesian Estimates for Terbinafine Trials

Arm	$\hat{\theta}_i^{NMM}$	$\hat{\theta}_i^{IMM}$	$\hat{\theta}_i^{MLE}$	$\hat{\theta}_i^{HB}$	Arm	$\hat{\theta}_i^{NMM}$	$\hat{\theta}_i^{IMM}$	$\hat{\theta}_i^{MLE}$	$\hat{\theta}_i^{HB}$
1	0.038	0.037	0.037	0.038	21	0.023	0.019	0.018	0.018
5	0.029	0.023	0.021	0.020	25	0.024	0.017	0.014	0.014
10	0.060	0.068	0.073	0.074	30	0.030	0.026	0.024	0.024
15	0.031	0.026	0.024	0.023	35	0.019	0.012	0.010	0.010
19	0.012	0.007	0.006	0.005	39	0.040	0.040	0.041	0.041

Of interest is comparison of interval widths for different empirical Bayesian techniques and the full hierarchical Bayesian analysis. Interval widths are given by

Table 2.2 Selected Empirical Bayesian and Hierarchical Bayesian Interval Widths for Terbinafine Trials

Arm	$\hat{\theta}_i^{NMM}$	$\hat{\theta}_i^{IMM}$	$\hat{\theta}_i^{MLE}$	$\hat{\theta}_i^{HB}$	Arm	$\hat{\theta}_i^{NMM}$	$\hat{\theta}_i^{IMM}$	$\hat{\theta}_i^{MLE}$	$\hat{\theta}_i^{HB}$
1	0.045	0.048	0.050	0.050	21	0.038	0.038	0.038	0.039
5	0.060	0.066	0.068	0.069	25	0.049	0.049	0.047	0.048
10	0.078	0.099	0.108	0.116	30	0.054	0.059	0.061	0.062
15	0.065	0.075	0.078	0.081	35	0.040	0.036	0.034	0.035
19	0.025	0.020	0.018	0.019	39	0.060	0.070	0.074	0.076

For observations at or near zero, changing the estimation method can noticeably affect the resulting empirical Bayesian and Bayesian estimates. In arm 19 of the study, for example, the interval width for the empirical Bayesian procedure with the naive moments estimator is roughly 30% larger than the corresponding hierarchical Bayesian estimator, and using the iterated method of moments still produces an interval that is roughly 5% larger.

The methods for addressing underestimation of the variance in Section 3.1 do not significantly change these results — the bias correction method as described in Carlin and Gelfand (1991) using the naive method of moments estimator produces an interval width of 0.055 (over twice as large) for arm 19, while the bootstrap method described in Laird and Louis (1987) using the naive method of moments estimator produces an interval width of 0.038 (roughly 52% larger). Since it works by addition of extra variance, Morris's method will only increase the width of the interval as well. It does not appear the idea that underestimation of variance is simply corrected by widening of the interval to approximate some vague Bayesian interval does not seem appropriate for this data set.

2.3.4 Explanation

So how did the empirical Bayesian estimator produce an interval width greater than the full hierarchical Bayesian procedure? Consider the posterior distributions for θ_{19} using empirical Bayes with naive method of moments estimates, iterated method of moments estimates, and maximum likelihood estimates

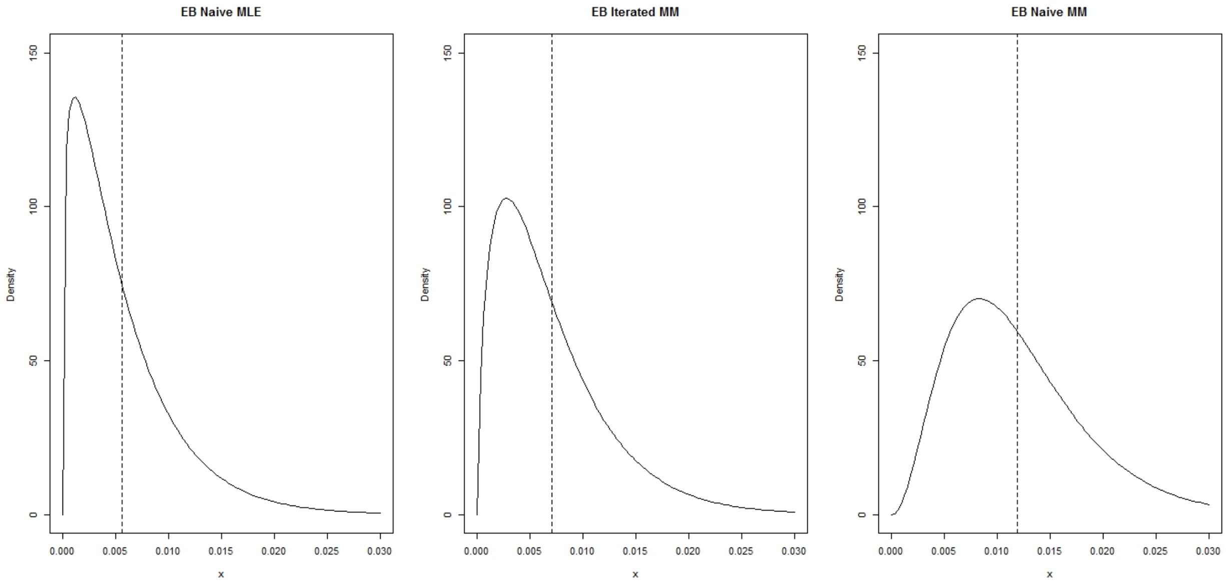

Figure 2.1 Empirical Bayesian Posterior Distributions for Adverse Reaction Proportion of Arm 19 of Terbinafine Trials

In the traditional normal model, a shift of location will not affect the width of a confidence interval. However, the doubly bounded nature of the of the beta distribution causes the shape of the posterior — and hence the width of the confidence interval — to vary dramatically as the mean of the posterior approaches zero or one. This can be seen on plot 2.1 above for the empirical Bayes posteriors for arm 19, with the dashed line representing the the expected value of the posterior, which is the empirical Bayesian estimate.

It would be possible to consider this issue as the result of differing estimation of the underlying variance parameter of the beta distribution M — the three estimates were $\hat{M} = 88.37$, $\hat{M} = 46.84$, and $\hat{M} = 36.01$ for the naive method of moments, iterated method of moments, and maximum likelihood respectively. Seeing such a large difference, it is no surprise that the empirical Bayesian estimates with the largest $\hat{M}$ estimate (and hence, smallest posterior variance) will have the smallest interval width. What is surprising is that, as shown with the method of moments empirical Bayesian interval for arm 19, the empirical Bayesian width may still be larger than even a full Bayesian analysis with weakly informative or noninformative hyperpriors. This phenomenon is not simply an issue of variances, however. Since the variance is also a function of the mean, it may occur even when the variance parameter M (and equivalently, the shrinkage coefficient B) is known, as will be shown directly.

2.3.5 Known M

For the sake of simplicity, first suppose that the the true data model is beta-binomial and the variance parameter M (or equivalently, the dispersion parameter ϕ) of the underlying beta distribution is known. For the Bayesian analysis, a hyperprior must be chosen for μ. Since μ must fall in the interval between zero and one, an obvious choice is the beta distribution

$$\mu \sim Beta(\lambda, M_0) \tag{2.12}$$

with λ and M_0 again representing the mean and variance parameter of the beta distribution. In following with the previously described framework of equations (2.11), a diffuse prior on μ is used

$$\mu \sim Beta(\hat{\mu}, 2) \tag{2.13}$$

Where $\hat{\mu}$ is the estimated mean, either by the method of maximum likelihood or the method of moments (equal sample sizes are assumed so that the naive and iterated method of moments produce equal estimates, though results are not dependent on this assumption). In this way, full hierarchical Bayes prior expectation is matched to moments of the the empirical Bayes prior distribution.

The prior in equation (2.13) uses $M_0 = 2$. Taking M_0 to ∞ will create prior that is degenerate, taking on $\hat{\mu}$ with probability 1 and producing the empirical Bayesian prior (and posterior), while taking M_0 to zero produces, as a limiting distribution, an alternative which will be referred to nominally as Haldane's prior.

$$p(\mu) \propto \frac{1}{\mu(1-\mu)} \tag{2.14}$$

Haldane's prior is an improper prior (in fact, it is an unnormalized $Beta(0,0)$ density); however, the posterior will be proper so long as $y_i \neq 0$ for all i and $y_i \neq n_i$ for all i.

When used as a straight prior for the θ_i in the binomial setting, Haldane's prior produces a posterior distribution with $E[\theta_i] = y_i/n_i$, the maximum likelihood estimate for θ_i. When used as a prior on μ in the hierarchical Bayesian setting with fixed M, it will produce intervals that match in what will be called the "Morris" sense — the hierarchical Bayesian posterior for θ_i will have the same expectation as (or very closely approximate the expectation of) the empirical Bayesian posterior with maximum likelihood estimator for the prior parameters

$$\hat{\theta}_i^{HB} \approx \hat{\theta}_i^{MLE} \tag{2.15}$$

In this way, hierarchical Bayesian intervals for an estimate $\hat{\theta}_i^{HB}$ may be used to quantify underestimation of variance with respect to maximum likelihood empirical Bayesian intervals around an estimate $\hat{\theta}_i^{MLE}$. In the beta-binomial case with known M, this will be true so long as the number of trials n_i and the number of observations k are both are not small (simulations suggest roughly three trials of three observations each). This approximation is discussed in Chapter 3.

A natural question arises: why not also attempt to construct a prior which gives posterior estimates equal to the empirical Bayesian estimates using the method of moments estimator? The answer is that since the Bayesian estimate depends strongly on the likelihood it will tend towards a maximum likelihood estimate, and as such it may not be possible to construct such a prior without making it at least moderately informative — and informative priors are not of great use in investigation of this problem. Note, however, that under the framework used in formula (2.12), Haldane's prior still exists as a limiting distribution no matter the estimator for μ.

2.3.6 Simulation Study

A simple simulation study was performed to assess the effect, simulating from the following model:

$$y_i \overset{indep}{\sim} Binomial(\theta_i, 10)$$

$$\theta_i \overset{iid}{\sim} Beta(0.10, 4)$$

for $i = 1, 2, ..., 10$ total observations. One hundred thousand Monte Carlo samples were taken for each of one thousand simulated data sets. Data sets such that $y_i = 0$ for all i or $y_i = n_i$ for all i were discarded. The variance parameter M (and correspondingly, the shrinkage coefficient B) was known and used in all estimation procedures. For each θ_i, the empirical Bayesian interval width was calculated using the method of moments estimator (since sample sizes are equal, the iterated and naive estimators will be equivalent) and marginal maximum likelihood. Full Bayesian analyses were also conducted using the beta priors given in Section 2.3.5 — the intermediate vague beta hierarchical distribution that matched the empirical Bayes prior expectation for both moment and maximum likelihood estimators (the prior in equation (2.13)), and Haldane's prior that matched to empirical Bayes posterior estimates using the maximum likelihood estimator (the prior in equation (2.14)).

Table 2.3 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.10, \text{Known } M = 4, n_i = k = 10$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}$	Haldane's Prior
Coverage	0.93	0.935	0.949	0.949	0.95
Width	0.233	0.236	0.241	0.241	0.241

The coverages here would be unconditional, using the definition in equation (2.4). Unsurprisingly, when the data simulated is directly from a beta-binomial model with known variance parameter M, both the empirical Bayesian and hierarchical Bayesian methods perform very well. The interval widths follow similarly — the average empirical Bayesian interval width is shorter than the average hierarchical Bayesian width, but not by much. The choice of prior does not appear to have a large influence on the posterior distribution in this particular example, as all three choices have nearly identical coverage and average widths, which should be unsurprising as Haldane's prior is simply the prior in equation (2.13), but fully diffused by taking $M_0 = 0$.

Of interest are the cases where the empirical Bayesian interval width is larger than the full hierarchical Bayesian interval width. Consider the distribution of the ratio of the empirical Bayesian interval widths using the method of moments estimator to full hierarchical Bayesian interval width using Haldane's prior, seen below in Figure 3.1.

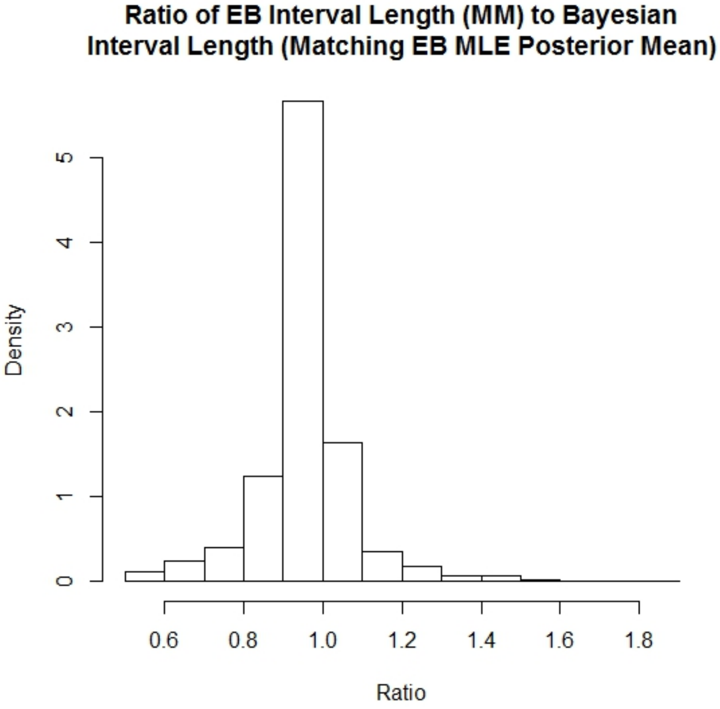

Figure 2.2 Ratio of Empirical Bayesian Interval Widths by Method of Moments to Hierarchical Bayesian Interval Widths by Haldane's Prior

In the simulated data set, 23.35% of empirical Bayesian intervals using the method of moments estimator are longer than the hierarchical Bayesian intervals. Conversely, the empirical Bayesian interval width using the maximum likelihood estimator is larger than the full hierarchical interval width in all but approximately 1% of cases for all three Bayesian priors. This may be attributed to simulation error from the MCMC procedure — that is, if the empirical

Bayesian and analytical Bayesian interval widths are very, very close, then there is a probability that the MCMC approximation to the Bayesian interval width will be larger than the empirical Bayesian interval width due simply to natural variation in the samples from the posterior distribution.

2.3.7 Analysis of Phenomenon

An interesting question exists as to when empirical Bayesian methods may, depending on the choice of estimation method and prior distribution, produce an interval that is longer than an interval produced by a full hierarchical Bayesian analysis (this will henceforth be referred to simply as "the phenomenon"). It should be stated beforehand that since the Bayesian analysis is driven by the likelihood, an empirical Bayesian analysis using a maximum likelihood estimator generally can not be larger than a full hierarchical Bayesian analysis except in an informative prior setting. For other estimators, the chances are dependent on the shape of the posterior distribution.

Suppose that y_i successes are observed in n_i trials for some observation i, still with known variance parameter M. In the traditional α, β form, the parameters of empirical Bayes posterior distribution are given as

$$\tilde{\alpha}_i = y_i + \hat{\mu}M$$
$$\tilde{\beta}_i = n_i - y_i + (1 - \hat{\mu})M$$

These may be converted back to to μ, M notation as

$$\tilde{\mu}_i = \frac{y_i + \hat{\mu}M}{y_i + \hat{\mu}M + n_i - y_i + (1 - \hat{\mu})M} = \frac{y_i + \hat{\mu}M}{n_i + M} \tag{2.16}$$

$$\tilde{M}_i = y_i + \hat{\mu}M + n_i - y_i + (1 - \hat{\mu})M = n_i + M \tag{2.17}$$

Taking the expected value, the empirical Bayesian estimator is $\hat{\theta}_i = \tilde{\mu}_i$. The empirical Bayesian interval, however, depends on more than just the estimator - an important feature is the shape of the two posterior distributions. Unlike the normal, the beta distribution is flexible in shape and bounded on either side - for example, if one of α or β (in the traditional parametrization) is larger than 1 *and* the other smaller than 1, then as Casella and Berger (2002) notes the resulting distribution will be J-shaped, not unimodal. Even if the distribution is unimodal, it need not be symmetric, as seen in Figure 2.1.

The form of the empirical Bayesian posteriors is a beta distribution. Empirical evidence suggests that the full Bayesian posterior using Haldane's prior (equation (2.14)) is very well approximated by a beta distribution with mean and variance

$$E[\theta_i] = \tilde{\mu} = \tilde{\mu}_i = \frac{y_i + \hat{\mu}_{MLE} M}{n_i + M} \tag{2.18}$$

$$Var(\theta_i) = \frac{\hat{\mu}_{MLE}(1-\hat{\mu}_{MLE})}{(M + n_i + 1)c} \tag{2.19}$$

and borrowing an approximation from Morris (1988), the constant c in equation (2.19) is given by

$$c = \frac{\tilde{\mu}_i(1-\tilde{\mu}_i)}{\tilde{\mu}_i(1-\tilde{\mu}_i) + B^2(n_i+M)\hat{\mu}_{MLE}(1-\hat{\mu}_{MLE})/[k(n_iB + (1-B)) + 1]} \tag{2.20}$$

where $B = M/(n_i + M)$. This approximation works extremely well for the simulated data set with $M = 4$ and $k = n_i = 10$. For different values of M, k, and n_i, simulations show the approximation appears to hold well, even when n_i is not constant for all i.

The approximation is further explored and justified in Chapter 3; however, a "conjugate-type relation" in the hierarchical beta-binomial model was noticed by Lee and Sabavala (1987) for a beta prior on the mean (and note that Haldane's prior is a beta distribution with parameters $\alpha = \beta = 0$), and Morris (1988) pretends the marginal distribution of the y_i given μ and M is a member of the natural exponential family in order to derive an exact conjugacy for the approximate distribution. In simulations, the relationship appears to be close to exact. Empirically, the root mean squared difference (over all θ_i estimates in all data sets) between the MLE and hierarchial Bayesian estimators for each of the θ_i is 0.000285, as compared to a 0.005944 (over 20 times larger) root mean squared difference between the MM and hierarchical Bayesian estimators for the θ_i. The idea of matching posterior estimates to frequentist estimates has a scattered history within statistics, though most have focused on the posterior mode rather than the posterior mean. A brief discussion of the history of matching posterior moments to frequentist estimates is provided in Ghosh and Liu (2011), along with examples of priors that, up to a high order of approximation, have posterior expectation matching the maximum likelihood estimate in the univariate and multivariate (but not hierarchical) case.

Unlike the normal density, the variance of the beta distribution is a function of the mean μ — for a fixed M, the variance is maximized at $\mu = 0.5$ and decreases as μ moves towards 0 or 1, and the distribution will become more compact. As this occurs, central intervals taken from that distribution will necessarily have a smaller width. Furthermore, the distribution is flexible and shape and the skew increases as the mean moves away from $\mu = 0.5$.

The phenomenon occurs when the empirical Bayesian posterior distribution based on an estimator $\hat{\mu}$ gains enough additional interval width from the increased posterior variance of $\hat{\mu}$ being larger than $\hat{\mu}_{MLE}$ and from the shape of the posterior becoming more symmetric as $\hat{\mu}$ moves away from the boundary of the parameter space towards 0.5 to overcome the additional variance gained from the full hierarchical model (for Haldane's prior the proportion increase is given by $\frac{1}{c}$ from equation (2.20), though the changing shape of the posteriors means that the required increase in $\hat{\mu}$ is different than the simple ratio of variances). The result is that there exists a value r such that if $\hat{\mu}/\hat{\mu}_{MLE} > r$, then the empirical Bayesian interval width will be larger than the full hierarchical Bayesian width. This can be seen in the pattern of Figure 2.3 below for $y_i = 0$.

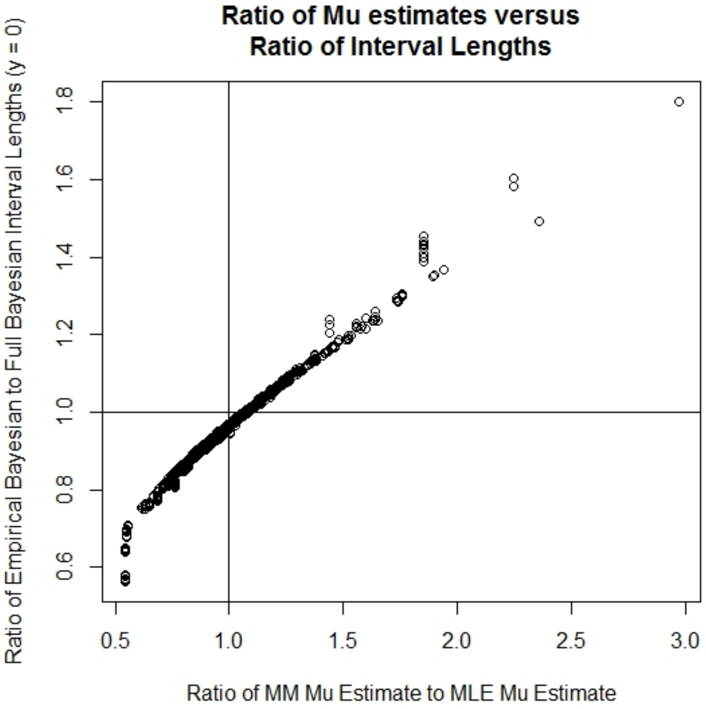

Figure 2.3 Ratio of Empirical Bayesian Interval Width using Method of Moments over Average (Within Data Set) Hierarchical Bayesian Interval width using Haldane's Prior versus Ratio of $\hat{\mu}$ Estimates (Method of Moments Over Maximum Likelihood) for $y_i = 0$

The phenomenon occurs when $\hat{\mu}_{MM}$ is about $r = 7.4\%$ larger than $\hat{\mu}_{MLE}$, though the exact value of r may vary slightly depending on the specific estimates. It can also be seen when $y_i \neq 0$. Supposing $y_i = 1$ shows a similar pattern in Figure 2.4 below.

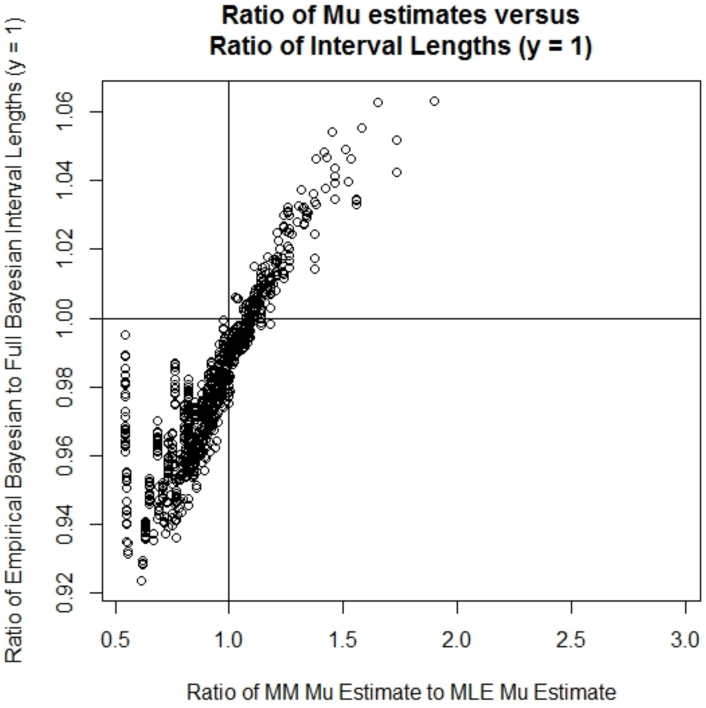

Figure 2.4 Ratio of Empirical Bayesian Interval Width using Method of Moments over Average (Within Data Set) Hierarchical Bayesian Interval width using Haldane's Prior versus Ratio of $\hat{\mu}$ Estimates (Method of Moments Over Maximum Likelihood) for $y_i = 1$

The phenomenon occurs when $\hat{\mu}_{MM}$ is about $r = 8.8\%$ larger than $\hat{\mu}_{MLE}$.

Similar results may be shown for any arbitrary y_i, though as y_i/n_i moves further away from 0 or 1, the effect (and the underestimation of the variance) decreases as posterior mean move closer to 0.5. Furthermore, though these simulations have assumed $\hat{\mu}_{MLE} < 0.5$, results will apply with inverted equalities when $\hat{\mu}_{MLE} > 0.5$.

2.3.8 Non-Bayesian Methods

It is useful to compare the coverage and average interval width for the simulated data of Section 2.3.6 to the non-Bayesian methods described in Section 2.1. The bias correction and parametric bootstrap methods were considered for both the method of moments (MM) and maximum likelihood estimators (MLE). Unconditional coverage and average interval width are given by

Table 2.4 Non-Bayesian Correction Methods Coverage and Interval Width — $\mu = 0.1, \text{Known M} = 4, k = n_i = 10$

	Bias Corrected MM	Bias Corrected MLE	Bootstrap MM	Bootstrap MLE
Coverage	0.959	0.957	0.956	0.956
Width	0.261	0.258	0.242	0.245

Despite bias correction method being intended to "correct" the interval width as opposed to simply widen it, the bias correction method actually performs the worst — the coverage is slightly above 95%, but at the cost of the interval width being, on average, even longer than the full Bayesian solution. The bootstrap method performs similarly to the full Bayesian analysis shown in Table 3.1, which is no surprise given that the bootstrap approximates a full Bayesian solution.

Further simulation studies were conducted in order to determine under what conditions the phenomenon is likely to occur and, in general, to compare interval width correction methods based on the beta-binomial model, with the result being that the phenomenon is most likely to appear when $\hat{\mu}$ is near 0 or 1 and M is small, corresponding to data sets with lots of group proportions near 0 or 1, as in the terbinafine data given in Appendix B. These results are described in Appendix C.

2.3.9 Unknown M

In most situations, the variance parameter M (or, equivalently, the dispersion parameter ϕ) must also be estimated. It is easiest to use the parametrization in equation (2.8) in order to specify priors on ϕ. Since both μ and ϕ are bounded by 0 and 1, beta distributions are obvious choices. In fitting with the previously discussed framework, priors that match the moments of the hyperpriors to the empirical Bayesian estimates are used.

$$
\begin{aligned}
\mu &\sim Beta(\hat{\mu}, 2) \\
\phi &\sim Beta(\hat{\phi}, 2)
\end{aligned}
\tag{2.21}
$$

In this way, each prior has expectation equal to the empirical Bayesian estimates, but sufficiently diffused. This may be used as an alternative to the Morris-style method of comparison, as the empirical Bayesian estimates using each method can be constructed as the result of a full hierarchical Bayesian analysis with the above priors placed on each parameter, but with the hyperprior variance parameter M_0 (equal to 2 in the cases above) going to infinity to create degenerate distributions that take the empirical Bayesian prior estimate with probability 1. The above prior will yield a proper posterior so long as $\hat{\mu}$ and $\hat{\phi}$ are not either zero or one, corresponding to extreme scenarios under which the beta-binomial model would not typically be considered.

In order to approximate the Morris-style comparisons, the following priors will be used

$$
\begin{aligned}
\mu &\propto \frac{1}{\mu(1-\mu)} \\
\phi &\sim Beta(0.5, 1)
\end{aligned}
\tag{2.22}
$$

The posterior expectation of the full beta-binomial model will very loosely approximate the empirical Bayesian estimate using the method of maximum likelihood, though the approximation is not nearly as close to exact as the Haldane's prior for the known M case. It will become very close as the sample sizes n_i and k increase, as shown in the results of the Terabifine analysis in Table 2.1. Simply taking $\mu \sim Beta(0,0)$ and $\phi \sim Beta(0,0)$ as in the case of known M generally does not approximate posterior means well for smaller M values, possibly due to unaccounted for covariance between $\hat{\mu}$ and $\hat{M}$ when $\hat{M}$ is small. Efforts to determine a weakly informative or or noninformative set of priors that produced posterior estimates of θ_i strongly approximating empirical Bayesian estimates using any estimator were unsuccessful, and it is unclear whether such priors exist. The set of priors in (2.22) produces Morris-style approximations for many different (but not all) data sets. As the prior on ϕ is proper, the posterior distribution is guaranteed to be proper so long as $y_i \neq 0$ for all i and $y_i \neq n_i$ for all i.

2.3.10 Further Simulations

A simulation study was performed, simulating data from a beta-binomial distribution with $\mu = 10$, $M = 4$, and $k = 10$ binomial observations of $n_i = 10$ trials each. Intervals compared were the method of moments and method of maximum likelihood empirical Bayesian methods and full Bayesian hierarchical analyses using the priors described in equations (2.21) (matching both methods of empirical Bayesian estimation) and the priors described in equations (2.22). Coverage and average interval widths are shown below in Table 2.5.

Table 2.5 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.10, M = 4, n_i = k = 10$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}, \hat{\phi}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}, \hat{\phi}_{MLE}$	Approx. Morris Matching Prior
Coverage	0.842	0.838	0.894	0.894	0.941
Width	0.221	0.220	0.229	0.229	0.245

The underlying problem with the use of empirical Bayesian analysis is much more apparent — using either estimation method, the empirical Bayesian intervals do not achieve nominal 95% coverage. Matching empirical Bayesian estimates to hierarchical prior means does not necessarily provide nominal coverage in this scenario; however, the noninformative prior that approximates Morris-style matching performs better, achieving almost nominal coverage.

The set of priors used in equation (2.22) does not match the empirical Bayesian results nearly as well as in the case with known M. The root mean squared difference between the MLE empirical Bayesian estimates for θ_i and approximate Morris matching Bayesian hierarchical estimates was 0.0467, as compared to 0.0471 using the MM empirical Bayesian estimates. Though they produced undercoverage, the intermediate priors of equation (2.21), however, had a root mean squared difference of 0.0231 for the MLE empirical Bayesian estimator and 0.0240 for the MM empirical Bayesian estimator.

Table 2.6 Non-Bayesian Correction Methods Coverage and Interval Width — $\mu = 0.1, M = 4, k = n_i = 10$

	Bootstrap	Bootstrap	Bias Correction	Bias Correction
	MM	MLE	MM	MLE
Coverage	0.939	0.917	0.944	1
Width	0.245	0.241	0.461	0.89

Non-Bayesian methods in this scenario provided mixed results. The bootstrap approach performs decently, with average coverage and width similar to the full hierarchical Bayesian analysis in the method of moments case, though there tends to be undercorrection in the case of the maximum likelihood estimator. The bias correction approach breaks down, though this is partly due to occasional poor estimates of the variance parameter M in the bootstrap procedure.

Further simulations were again conducted in order to determine the appearance of the phenomenon and to determine appropriate priors for comparison to the empirical Bayesian intervals — in general, the bias correction technique fails when the posterior distribution is not sufficiently normal, while the bootstrap performs well. For a large M value, the intermediate priors of equation (2.21) show near nominal coverage. These results are described in Appendix D.

2.3.11 Recommendations

For the case of known variance parameter M, the empirical Bayesian interval may be compared to a Bayesian interval produced by using a diffuse prior with mean $\hat{\mu}$ — the particular choice of estimator does not appear to matter. By fully diffusing to Haldane's prior, however, empirical Bayesian interval widths produced from the method of maximum likelihood can be directly compared in the Morris style. Even without using Haldane's prior, however, a general noninformative prior on μ should produce excellent results. For the method of moments, the Bayesian width using Haldane's prior may still be useful for comparison if one is not concerned about matching posterior expectations, and may indicate intervals that are inappropriately centered and, hence, too long. If one is concerned about matching posterior expectations, the bootstrap correction method appears to offer the best coverage.

For the case of unknown variance parameter M, the situation is not as clear, as there is no known prior that necessarily matches posterior Bayesian estimates to empirical Bayesian estimates for any method of estimation. Depending on the data set, using Haldane's prior on μ and a Beta(0.5,1) prior on ϕ or beta priors that match means of μ and θ to empirical Bayesian estimates may tend to provide nominal coverage, and can be used for comparison. In all cases, the parametric bootstrap of Laird and Louis (1987) will produce useful intervals that match the posterior expectation, though coverage may be slightly more or less than the nominal amount, and interval widths may be longer than a corresponding Bayesian solution that produces nominal coverage.

2.4 Future Work

Further work on this project is possible. Several theoretical results have been stated with little justification, though the works of Carl Morris (particularly Morris (1988)) provide some evidence for the approximations used in this paper, and are further explored in Chapter 3. Prior selection for the case of unknown variance parameter M remains undeveloped. Simulations suggest that simply taking diffuse beta priors on μ and ϕ with expectation equal to $\hat{\mu}$ and $\hat{\phi}$ does not produce nominal coverage for small M values, corresponding to a large spread of θ_i values, though as the data-generating M increases and the spread of θ_i decreases (approaching a simple binomial model with fixed probability of success θ) these intermediate priors perform increasingly well.

2.5 Conclusion

In general, the idea of underestimation of the variance is well-defined in the normal-normal case. Stepping outside of that particular model, however, ideas do not apply as neatly, as for a given set of empirical Bayesian estimates $\hat{\theta}_i$, there may not exist a hierarchical Bayesian prior which produces posterior estimates for the θ_i equivalent to the empirical Bayesian estimates. Even if there exists a prior which offers a close approximation, the existence of multiple estimation techniques makes it entirely possible to have an empirical Bayesian interval based upon some consistent estimator which produces an interval width larger than the hierarchical Bayesian interval. The types of data sets where these phenomenon may occur — small sample data sets, with observations tending near zero or one — are often the types where Bayesian and empirical Bayesian techniques are the most useful, and the difference in interval width may be non-trivial.

Non-Bayesian correction methods have issues as well. The bias correction method intended to "correct," rather than simply widen, appears to perform poorly in the small sample beta-binomial case, and while the parametric bootstrap correction method is superior at producing interval estimates which achieve nominal unconditional coverage, in its basic form it is artificially creating a prior around an estimate rather than matching a known hierarchical prior. There is no clear reason why this should be considered "correct," especially if there exists a hierarchical prior that produces near-nominal coverage and a shorter interval width.

An alternative method of comparison is given by an intermediate vague prior with mean given by estimates of prior parameters. This method focuses less on the idea of correcting empirical Bayesian intervals by matching empirical Bayesian estimates exactly to expectations of hierarchical Bayesian models, but does provide a framework under which the hierarchical prior used in Morris-style corrections for the normal-normal model can be shown as a limit of the intermediate prior density, and a similarly limiting density under the beta-binomial model matches and can be used to quantify the underestimation in variance in when using the maximum likelihood estimator and provide an appropriate width correction with other estimators. In the case of known shrinkage parameter B, this provides excellent nominal coverage. In the case of unknown shrinkage parameter, the method generally produces good (though not necessarily nominal) coverage in many cases and often outperforms non-Bayesian correction methods, though further analysis is needed.

CHAPTER 3. COMPARISON FRAMEWORK FOR THE NATURAL EXPONENTIAL FAMILY WITH QUADRATIC VARIANCE FUNCTIONS

3.1 NEFQVF Families

The results of Chapter 2 may be extended to a larger class of models, of which the beta-binomial is a member - the natural exponential family with quadratic variance function (NEFQVF). In this chapter, properties of the NEFQFV family will be identified as relating to empirical Bayesian analysis, and some approximations for posterior distributions will be given and discussed, along with a framework for comparisons of widths of empirical Bayesian intervals, with a specific further application of the results to the gamma-Poisson model. These approximations can be used to refine a method from Morris (1988) for correcting empirical Bayesian interval widths with known shrinkage amount, and simulations for this method using the beta-binomial and gamma-Poisson models show near-nominal coverage.

3.1.1 NEFQVF Distributions

Properties of NEF and NEFQVF distributions have been studied and enumerated in Morris (1982), Morris (1983a), and Morris and Lock (2009), which define and give basic properties of NEFQVF distributions, combine and present statistical theory for NEFQVF distributions, and show connections among NEF and NEFQVF distributions, respectively. Sections 3.1.1 and 3.1.2 will draw and intermix results from these papers freely. Members of the NEFQVF family are, as the name implies, members of the natural exponential family, and hence have densities that may be written proportional to

$$\exp(x\zeta - \psi(\zeta))dF(x)$$

where ζ is the natural parameter, and $E[x] = \psi'(\zeta) = \theta$. Each NEF has a variance that is a function of the mean, $Var(X) = \psi''(\zeta) = V(\theta)$. The "quadratic variance function" refers to the fact that for each of these distributions, the variance is a polynomial function of the mean with degree two or less.

$$V(\theta) = v_0 + v_1\theta + v_2\theta^2 \tag{3.1}$$

The six NEF distributions with QVF are the normal (with known variance σ^2), the binomial, the Poisson, the gamma, the negative binomial, and the natural exponential family generated as a convolution of hyperbolic secant distributions.

3.1.2 NEFQVF Shrinkage

Each of the NEFQVF distributions may be written as the convolution of "generator" distributions - the normal is the sum of normals, the Poisson is the sum of Poissons, the binomial is the sum of Bernoullis, the gamma is the sum of exponentials, the negative binomial is the sum of geometrics, and the sixth NEFQVF is, as the name implies, a convolution of hyperbolic secant distributions.

To construct a framework for shrinkage estimation for NEFQVF distributions, begin by defining $x_{i,j}$ as independent observation j from an unspecified NEFQVF family with given expected value θ_i (and corresponding natural parameter ζ_i), where $j = 1, 2, ..., n_i$ and $i = 1, 2, ..., k$.

$$x_{i,j}|\theta_i \overset{iid}{\sim} NEFQVF[\theta_i, V(\theta_i)]$$

In general, the notation $[\bullet, \bullet]$ will be taken as a distribution with specified mean and variance but unspecified form, with a possible identifier in front to indicate family (NEF to mean natural exponential family, for example, with QVF indicating quadratic variance function).

Since the sum of NEFQVF observations is also NEFQVF, define

$$y_i = \sum_{j=1}^{n_i} x_{i,j}$$

These y_i will will have mean and variance

$$y_i|\theta_i, n_i \overset{indep}{\sim} NEFQVF\left[n_i\theta_i, n_i V(\theta_i)\right]$$

and by extension

$$\frac{y_i}{n_i}|\theta_i, n_i \overset{indep}{\sim} NEF\left[\theta_i, \frac{V(\theta_i)}{n_i}\right]$$

As members of the natural exponential family, each of the NEFQVF distributions are guaranteed to have conjugate distributions, and the NEFQVF family in particular features many well-known examples, such as the normal prior for the normal mean, the beta prior for the binomial success probability, the gamma prior for the Poisson rate parameter, and the inverse gamma prior for the exponential rate parameter.

For the conjugate distributions, the natural parameter ζ_i for each group is assumed to have density

$$\zeta_i \overset{iid}{\sim} K(\mu, M)\exp(\mu M\zeta_i - M\Psi(\zeta_i))$$

where $K(\mu, M)$ is a normalizing constant and $E[\Psi'(\zeta_i)] = E[\theta_i] = \mu$ is the expected value of the θ_i. This also yields a density on θ_i given by

$$g(\theta_i|\mu, M) = \frac{K(\mu, M)\exp\{M\mu\kappa(\theta_i) - M\Psi(\kappa(\theta_i))\}}{V(\theta_i)} \tag{3.2}$$

where $\kappa(\theta_i)$ is the inverse function of $\theta_i = \Psi'(\zeta_i)$ and $V(\theta_i)$ is the variance function of the original $x_{i,j}$ applied to the prior means θ_i, with v_2 as in equation (3.1). This prior distribution has first two moments

$$\theta_i|\mu, M \overset{iid}{\sim} CD\left[\mu, \frac{V(\mu)}{M - v_2}\right] \tag{3.3}$$

where $V(\mu)$ is again the original variance function of the $x_{i,j}$ applied to the prior mean μ. Letting the θ_i follow the conjugate distribution with density given in equation (3.2), the posterior distribution of the θ_i is then also the conjugate distribution (CD) with mean and variance

$$\theta_i|y_i, \mu, M, n_i \sim CD\left[(1-B)\frac{y_i}{n_i} + B\mu, \frac{V\left((1-B)\frac{y_i}{n_i} + B\mu\right)}{n_i + M - v_2}\right] \tag{3.4}$$

where the quantity B is the shrinkage parameter, which for members of the natural exponential family has the form

$$B = \frac{E[V(\theta_i)]}{E[V(\theta_i)] + n_i Var(\theta_i)} = \frac{M}{n_i + M}$$

with expectation and variance taken with respect to the conjugate density $g(\theta_i)$. The marginal distribution of the y_i/n_i integrating out over the θ_i is

$$\frac{y_i}{n_i}|\mu, M \sim \left[\mu, \frac{V(\mu)}{n_i B - v_2(1-B)}\right] \tag{3.5}$$

This marginal distribution may or may not be NEF.

3.1.3 Morris Hyperpriors

In Morris (1988), Carl Morris uses a similar setup, additionally assuming instead that $n_i = n$ for all i (the necessity of this assumption will be discussed at the end of this section), and notes that if the marginal distribution in equation (3.5) is simply assumed to be NEF with the same two moments, then a naive estimator of the prior mean μ may be constructed as

$$\hat{\mu} = \frac{1}{k}\sum_{i=1}^{k} \frac{y_i}{n} \tag{3.6}$$

This estimator is complete and sufficient for μ with mean and variance

$$\hat{\mu} \sim NEF\left[\mu, \frac{V(\mu)}{k(nB - v_2(1-B))}\right]$$

Note that the marginal NEF assumption is not entirely unreasonable - the marginal distributions of equation (3.5) are mixtures of NEFQVF distributions, and in the normal-normal cases is exactly NEF. This assumption treats, for example, the beta-binomial distribution as if it were a binomial.

Suppose B (or equivalently M) is assumed known and a hyperprior is placed on μ given by

$$h(\mu) \propto \frac{1}{V(\mu)} \tag{3.7}$$

which is the variance function of the original $x_{i,j}$ applied to the mean μ of the prior distribution. This hyperprior will nominally be called the Morris hyperprior. Morris (1988) identifies that "this latter prior density is equivalent to a conjugate prior distribution." Then using the conjugacy property of NEF distributions applying equation (3.4) with $0, k(nB - v_2(1-B))$, and $\hat{\mu} = \frac{1}{k}\sum_{i=1}^{n} y_i/n$ taking the place of M, n, and y_i/n, the posterior mean and variance are

$$\mu|\hat{\mu}, B \sim \left[\hat{\mu}, \frac{V(\hat{\mu})}{k(nB - v_2(1-B)) - v_2}\right] \tag{3.8}$$

and the posterior distribution of θ_i is given by

$$\theta_i|y_i, n, B \sim \left[\tilde{\mu}_i = (1-B)\frac{y_i}{n} + B\hat{\mu}, \tilde{s}_i^2\right] \tag{3.9}$$

where

$$\tilde{s}_i^2 = \frac{V(\tilde{\mu}_i)}{n + M - v_2} + B^2\left(\frac{n+M}{n+M-v_2}\right)\frac{V(\hat{\mu})}{k(nB - v_2(1-B)) - v_2} \tag{3.10}$$

Though Morris does not specify the distributional forms of equations (3.8) and (3.9), under the marginal NEF assumption the established conjugacy of the prior given in equation (3.7) yields that equation (3.8) will be of the same assumed conjugate distribution as equations (3.2) and (3.4). Empirical evidence furthermore suggests that the posterior distributions for θ_i using the Morris hyperprior will then also be of the same conjugate distribution as equations (3.2) and (3.4). This extends the conjugate form to a limited hierarchical setting.

The approximation makes the assumption that the distribution of equation (3.8) is NEF, which not necessarily true in most cases, but the necessity of this assumption appears to be weak for the cases considered. In the normal-normal model, the marginal distribution of the y_i given μ and M is normal, the conjugate prior on θ_i is normal, and the posterior distributions are normal, so the distribution in equation (3.9) is normal as well. Corrected intervals based on the normal-normal hierarchical model with the Morris hyperprior $h(\mu) \propto 1$ can be found in Morris (1983b), which also corrects for an unknown B (these are closely related to the James-Stein estimator). In the case of the gamma-Poisson model, explored in Section 3.2.1, the marginal distribution of the y_i given μ and M is negative binomial — not jut NEF, but NEFQVF — though in the μ and M parametrization, the maximum likelihood estimator $\hat{\mu}_{MLE}$ is not $\frac{1}{k}\sum_{i=1}^{k} y_i/n_i$. In the case of the beta-binomial model, the marginal distribution of the y_i given μ and M follows the beta-binomial distribution, which is not NEF. Empirically, assuming a conjugate distribution for the posterior distribution of equation (3.9) appears to work very well for the gamma-Poisson, as will be shown in Section 3.3, and beta-Binomial models, as was shown in Chapter 2, and is exact for the normal-normal.

Furthermore, the expected value of the full hierarchical posterior in equation (3.9) empirically appears to be much more closely approximated by the maximum likelihood estimator obtained by maximizing the log-likelihood using the marginal distribution of equation (3.5) with known M rather than the moment based estimator of equation (3.6). This should come as no surprise - in one-parameter natural exponential families, the posterior mean using the conjugate prior is a combination of the prior mean and the maximum likelihood estimate of the parameter. What is surprising is the strength of the approximation in the assumed NEF scenario. In the case of the normal-normal model, the moment and likelihood estimators are both given as equation (3.6), apart from positive-part corrections. In the beta-binomial model, as discussed in the previous chapter, both appear to be very close to each other, while the gamma-Poisson model will be discussed in Section 3.2.1. The remaining three NEFQVF distributions remain unstudied.

Utilizing these approximations, the full posterior distribution of the θ_i is estimated by a conjugate distribution of the same form as equation (3.4) (beta for binomials, normals for normals, gammas for Poissons, and so on) with mean and variance given by equations (3.9) and (3.10), but with $\hat{\mu}_{MLE}$ in place of $\hat{\mu} = \frac{1}{k}\sum_{i=1}^{n} y_i/n_i$. Simulations suggests that this approximation generally works well for different value of M, n, and k. Furthermore, Morris assumes $n_i = n$ for all n. Simulations also suggest that this is an unnecessary assumption, and using the maximum likelihood estimator $\hat{\mu}_{MLE}$ in place of the naive estimator of equation (3.6) appears to account for the differences in n_i between samples. This is shown in the coverage and interval width simulations of Section 3.3.

3.2 Interval Comparison Framework

The comparison framework assumes the NEFQVF model of Section 3.1.2 with known variance parameter M but unknown mean μ of the second-stage conjugate density given in equation (3.2).

A hyperprior of the same conjugate distribution as equation (3.2) is placed on the second-stage mean μ - a normal hyperprior is placed upon a normal mean, a beta hyperprior is placed upon a beta mean, a gamma hyperprior is placed upon a gamma mean, etc. - with mean λ and variance parameter M_0 chosen by the statistician. In doing so, the form of the hyperprior on μ is given by

$$h(\mu|\lambda, M_0) = \frac{K(\lambda, M_0)\exp(M_0\lambda\kappa(\mu) - M_0\Psi(\kappa(\mu)))}{V(\mu)} \tag{3.11}$$

Setting $\lambda = \hat{\mu}$, where $\hat{\mu}$ is some estimator, produces a model such that both the empirical bayes prior and full hierarchical prior (integrating out over μ) have expectation $\hat{\mu}$, but the full hierarchical prior has expanded variance. This prior is intermediate in the sense that taking M_0 to 0 (while ignoring the normalizing constant K) in equation (3.11) produces the Morris hyperprior of equation (3.7), while taking M_0 to ∞ in the moments of equation (3.3) (with M_0 in place of M and $\lambda = \hat{\mu}$ in place of μ) produces a degenerate hyperprior that takes on $\hat{\mu}$ with 0 variance (and probability 1), producing the empirical Bayesian prior. For the purposes of comparing empirical Bayesian intervals, this allows the intervals produced by a full Bayesian analysis using the Morris hyperprior to be used as a baseline for comparison.

From this, it can occasionally be seen that an empirical Bayesian interval is wider than the correct interval for that data set. Suppose empirical Bayesian intervals are obtained based off of some consistent estimator $\hat{\mu}$ where $\hat{\mu} \neq \hat{\mu}_{MLE}$ — the empirical Bayesian interval width based on $\hat{\mu}$ will occur when $\hat{\mu}$ is far enough away from $\hat{\mu}_{MLE}$ to overcome the increased width from the additional variance introduced by the full hierarchical Bayesian analysis using the Morris hyperprior and the changing shape of the posterior distribution (since the posteriors will not be perfectly symmetric, except in the case of the normal-normal model). This is not a simple ratio of variances (and in fact, the ratio of variances may be off by a significant amount), but the nature of the conjugate distribution for the empirical Bayes posterior and approximate conjugate distribution for the hierarchical Bayes posterior centered at $\hat{\mu}_{MLE}$ but with an increased variance implies that there exists a value r such that if $\hat{\mu}/\hat{\mu}_{MLE}$ is approximately greater than or less than r (depending on the variance function $V(\mu)$) then the empirical Bayesian interval based on $\hat{\mu}$ will be wider than the interval based full hierarchical Bayesian posterior using the Morris hyperprior. It can be said that as M increases, the required distance increases as to become nigh impossible, and as k and n_i increase, $\hat{\mu}$ and $\hat{\mu}_{MLE}$ should converge and the additional variance from the hierarchical Bayes estimate will be trivial.

3.2.1 The Gamma-Poisson Model

These properties may be illustrated by considering the Gamma-Poisson model. Let observation $x_{i,j}$ be independent and identically distributed observations from a Poisson distribution with a common mean θ_i

$$x_{i,j}|\theta_i \overset{iid}{\sim} Poisson(\theta_i)$$

which has variance function $V(\theta_i) = \theta_i$, with $v_0 = v_2 = 0$ and $v_1 = 1$. Let the θ_i be independent observations from a Gamma distribution

$$\theta_i \overset{indep}{\sim} Gamma(\mu, M) \tag{3.12}$$

The density of the the Gamma distribution in equation (3.12) is given by

$$f(\theta_i|\mu, M) = \frac{M^{\mu M}}{\Gamma(\mu M)} \theta_i^{\mu M - 1} e^{-M\theta_i}$$

In this parametrization, $E[\theta_i] = \mu$ and $Var(\theta_i) = \mu/M$ (in the traditional α, β parametrization, these would be $\mu = \alpha/\beta$ and $M = \beta$). This gives a marginal distribution of y_i that is negative binomial with mass function

$$p(y_i|n_i,\mu,m) = \frac{M^{\mu M} n_i^{y_i}}{\Gamma(\mu M) y_i!} \frac{\Gamma(y_i+\mu M)}{(n_i+M)^{y_i+\mu M}}$$

and dropping terms that do not involve the prior parameters μ and M, the log-likelihood is

$$\ell(\mu,M) = k\mu M \log(M) - k\log(\Gamma(\mu M)) + \sum_{i=1}^{k} [\log(\Gamma(y_i+\mu M)) - (y_i+\mu M)\log(n_i+M)] \tag{3.13}$$

For equal sample sizes n_i, the method of moments estimator $\hat{\mu}_{MM}$ is given by the naive estimator Morris uses in equation (3.6). The maximum likelihood estimator $\hat{\mu}_{MLE}$ is obtained by maximizing equation (3.13) with a fixed M. Using either method, the empirical Bayesian posterior distribution for θ_i has a gamma density with mean and variance

$$\theta_i|y_i,\mu,M,n_i \sim Gamma\left[(1-B)\frac{y_i}{n_i} + B\hat{\mu}, \frac{V\left((1-B)\frac{y_i}{n_i} + B\hat{\mu}\right)}{n_i+M}\right] \tag{3.14}$$

where $B = M/(M+n_i)$.

The Morris hyperprior on μ is given by

$$h(\mu) \propto \frac{1}{V(\mu)} = \frac{1}{\mu} \tag{3.15}$$

The posterior expectation $\tilde{\mu}_i$ using this hyperprior will closely approximate the expectation of equation (3.14) when the maximum likelihood estimator $\hat{\mu}_{MLE}$ is used. Using this hyperprior, the posterior distribution given in equation (3.9) is

$$\theta_i|y_i,n,B \sim Gamma\left[\tilde{\mu}_i = (1-B)\frac{y_i}{n_i} + B\hat{\mu}_{MLE}, \tilde{s}_i^2 = \frac{\tilde{\mu}_i}{n_i+M} + B^2\frac{\hat{\mu}_{MLE}}{k(n_iB)}\right] \tag{3.16}$$

3.2.2 Interval Width Comparison

Following the development of Section 3.1.3, the following prior distributions were considered for a full hierarchical Bayesian analysis

$$\mu \sim Gamma(\lambda, M_0) \tag{3.17}$$

which has density given by

$$h(\mu) = \frac{M_0^{M_0\lambda}}{\Gamma(M_0\lambda)} \mu^{M_0\lambda - 1} e^{-M_0\mu}$$

Setting $\lambda = \hat{\mu}$ produces the intermediate prior distribution of equation (3.11). Taking M_0 to ∞ produces a degenerate distribution that takes on $\hat{\mu}$ with probability 1, and taking M_0 to 0 (and ignoring the normalizing constant) gives

$$h(\mu) \propto \frac{1}{\mu} \tag{3.18}$$

which is the Morris hyperprior for μ.

3.2.3 Simulated Data

To test this, data was simulated from the model in Section 3.2.1 with $\mu = 0.1$, $M = 4$, and $n_i = 10$ for all of $k = 10$ groups. A total of one hundred thousand MCMC draws were taken for each of one thousand data sets.

For each data set, empirical Bayesian interval estimates were calculated using $\hat{\mu}_{MM}$ and $\hat{\mu}_{MLE}$, and three hierarchical Bayesian analyses were fit - one using the prior in equation (3.17) with $\lambda = \hat{\mu}_{MM}$ and $M_0 = 2$, one using the prior in equation (3.17) with $\lambda = \hat{\mu}_{MLE}$ and $M_0 = 2$, and one using the Morris hyperprior distribution given in equation (3.18). Data sets consisting of $y_i = 0$ for all i were discarded. Coverage and average width of 95% central intervals are shown below in Table 3.1.

Table 3.1 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.10, \text{Known } M = 4, n_i = k = 10$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}$	Morris Prior
Coverage	0.925	0.936	0.950	0.950	0.950
Width	0.263	0.266	0.271	0.271	0.271

Coverage is superior with the intermediate and Morris hyperprior, both achieving slightly above nominal coverage, and the empirical Bayesian intervals using a maximum likelihood estimator are superior to intervals using the method of moments estimator. In terms of approximation, the root mean squared difference between the empirical Bayesian estimates for θ_i using the maximum likelihood estimator $\hat{\mu}_{MLE}$ and the hierarchical Bayesian estimates using the Morris hyperprior was 0.00051, as opposed to 0.00928 when using the method of moments estimator $\hat{\mu}_{MM}$ (equivalent to the naive estimator of equation 3.6) and the hierarchical estimates with the Morris hyperprior (roughly 18 times larger).

Furthermore, there exists a ratio r such that if $\hat{\mu}_{MM}/\hat{\mu}_{MLE} > r$, the empirical Bayesian width using the method of moments estimator exceeds the width of hierarchical Bayesian intervals using the Morris hyperprior. This is shown below for $y_i = 0$ in Table 3.1

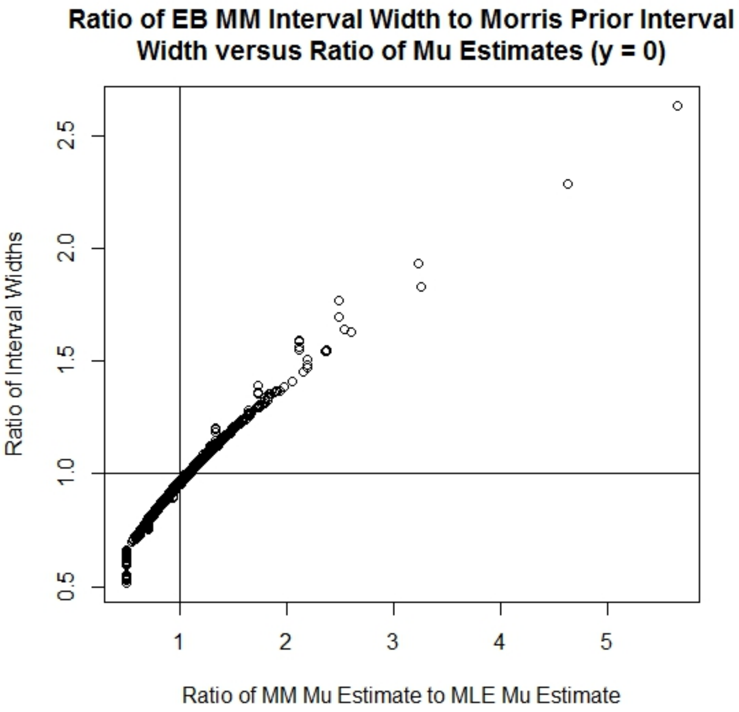

Figure 3.1 Ratio of Method of Moments Empirical Bayesian Interval Width to Morris Hyperprior Bayesian Interval Width by Ratio of $\hat{\mu}_{MM}$ over $\hat{\mu}_{MLE}$, $y_i = 0$

For this particular y_i, μ, M, n_i, and k, the ratio is about 1.07 — if $\hat{\mu}_{MM}$ is about 7% larger than $\hat{\mu}_{MLE}$, the ratio of empirical to full Bayesian interval widths will be larger than 1. This occured 27.75% of the time in the simulated data sets. Similar ratios exist when $y_i \neq 0$.

3.3 Corrected Intervals

The results of Section 3.1.3 can also be used to provide a correction to the empirical Bayesian interval with known shrinkage parameter B. Morris (1988) suggests using the moments from the posterior in equation (3.9) and constructing an interval as $\tilde{\mu} \pm 1.96\tilde{s}_i$. This technique works well for the normal-normal model, but for distributions that may be skewed, a large enough sample size for the posterior distribution to be well-approximated by a normal is likely to be a large enough sample size so that the uncertainty in $\hat{\mu}$ is small and does not neccesarily need to be accounted for.

A better interval estimator is constructed by using the mean and variance of the approximate conjugate distribution in equation (3.9) to solve for the two parameters and quantiles may be taken from the posterior. Using the gamma-Poisson model of Section 3.2.1, the Morris corrected empirical Bayesian posterior distribution is given in equation (3.16). Using $\tilde{\mu}_i$ and $\tilde{s}_i^2$ as the mean and variance, respectively, the traditional α, β parametrization is then

$$\begin{aligned} \tilde{\beta}_i &= \frac{\tilde{\mu}_i}{\tilde{s}_i^2} \\ \tilde{\alpha}_i &= \tilde{\mu}_i \tilde{\beta}_i \end{aligned} \tag{3.19}$$

Quantiles for can then be taken from a $Gamma(\tilde{\alpha}_i, \tilde{\beta}_i)$ to form central or highest posterior density credible intervals for θ_i.

Simulation results are shown below for the gamma-Poisson model. For each set of generating conditions, ten thousand data sets were simulated and central 95% intervals were taken from the the empirical Bayesian posterior distribution using a maximum likelihood estimator $\hat{\mu}_{MLE}$ and the Morris corrected posterior distribution of equation (3.19). Data sets such that $y_i = 0$ for all i were discarded. Coverage and average interval width are shown for the empirical Bayesian intervals (left) and Morris corrected empirical Bayesian intervals (right) below in Table 3.2.

Table 3.2 Naive MLE EB (Left) and Morris Corrected EB (Right) Interval Widths and Coverages for the Gamma-Poisson Model

Generating Conditions	Interval Coverage	Average Interval Width
$\mu = 0.1, M = 4, n_i = 10, k = 10$	0.933, 0.943	0.267, 0.275
$\mu = 0.1, M = 4, n_i = 30, k = 30$	0.948, 0.950	0.167, 0.167
$\mu = 0.1, M = 4, n_i = 3, k = 3$	0.845, 0.934	0.616, 0.746
$\mu = 0.1, M = 4, n_i = 100, k = 10$	0.948, 0.948	0.270, 0.270
$\mu = 0.1, M = 4, n_i = 10, k = 100$	0.947, 0.951	0.094, 0.094
$\mu = 0.1, M = 100, n_i = 10, k = 10$	0.828, 0.944	0.115, 0.162
$\mu = 0.1, M = 100, n_i = 100, k = 10$	0.939, 0.951	0.087, 0.091
$\mu = 0.1, M = 4, n_i \approx 10, k = 10$	0.934, 0.945	0.272, 0.281

For the last row, the n_i were drawn from a Poisson(10) distribution for each data set, with $n_i < 3$ discarded. The Morris corrected empirical Bayesian intervals generally succeed in obtaining slightly below nominal coverage in most scenarios, even for very small sample sizes ($n_i = k = 3$) and situations where the empirical Bayesian intervals drastically undercover ($M = 100$ and $n_i = k = 10$).

For the beta-binomial model, the variance function is given by $V(\theta_i) = \theta_i(1-\theta_i) = \theta_i - \theta_i^2$, so $v_0 = 0, v_1 = 1$, and $v_2 = -1$. Then assuming $B = M/(M+n_i)$ is known and using the Morris hyperprior of $h(\mu) = 1/[\mu(1-\mu)]$, applying equation (3.9) and assuming a beta form gives

$$\theta_i|y_i, n_i, M \sim Beta\Big[\tilde{\mu}_i = (1-B)\frac{y_i}{n} + B\hat{\mu}_{MLE}, \\ \tilde{s}_i^2 = \frac{\tilde{\mu}_i(1-\tilde{\mu}_i)}{n_i+M+1} + B^2\left(\frac{n_i+M}{n_i+M+1}\right)\frac{\hat{\mu}_{MLE}(1-\hat{\mu}_{MLE})}{k(n_iB+(1-B))+1}\Big]$$

This mean and variance can be converted to the traditional α, β parametrization by

$$\begin{aligned} \tilde{\alpha}_i &= \tilde{\mu}_i\left(\frac{\tilde{\mu}_i(1-\tilde{\mu}_i) - \tilde{s}_i^2}{\tilde{s}_i^2}\right) \\ \tilde{\beta}_i &= (1-\tilde{\mu}_i)\left(\frac{\tilde{\mu}_i(1-\tilde{\mu}_i) - \tilde{s}_i^2}{\tilde{s}_i^2}\right) \end{aligned} \tag{3.20}$$

and quantiles may be taken from a beta distribution with these parameters to construct intervals.

Simulation results are shown below for the beta-binomial model. For each set of generating conditions, ten thousand data sets were simulated and central 95% intervals were taken from the the empirical Bayesian posterior distribution using a maximum likelihood estimator $\hat{\mu}_{MLE}$ and the Morris corrected posterior distribution of equation (3.20). Data sets such that $y_i = 0$ for all i or $y_i = n_i$ for all i were discarded. Coverage and average interval width are shown for the empirical Bayesian intervals (left) and Morris corrected empirical Bayes intervals (right) below in Table 3.3.

Table 3.3 Naive MLE EB (Left) and Morris Corrected EB (Right) Interval Widths and Coverages for the Beta-Binomial Model

Generating Conditions	Interval Coverage	Average Interval Width
$\mu = 0.1, M = 4, n_i = 10, k = 10$	0.937, 0.945	0.239, 0.244
$\mu = 0.1, M = 4, n_i = 30, k = 30$	0.947, 0.949	0.152, 0.153
$\mu = 0.1, M = 4, n_i = 3, k = 3$	0.875, 0.942	0.469, 0.542
$\mu = 0.1, M = 4, n_i = 100, k = 10$	0.947, 0.950	0.086, 0.086
$\mu = 0.1, M = 4, n_i = 10, k = 100$	0.949, 0.950	0.243, 0.244
$\mu = 0.1, M = 100, n_i = 10, k = 10$	0.827, 0.945	0.109, 0.152
$\mu = 0.1, M = 100, n_i = 100, k = 10$	0.939, 0.951	0.082, 0.086
$\mu = 0.1, M = 100, n_i = 10, k = 100$	0.938, 0.950	0.110, 0.116
$\mu = 0.1, M = 4, n_i \approx 10, k = 10$	0.935, 0.943	0.244, 0.249

Again, the last row has n_i drawn from a Poisson(10) distribution for each data set with $n_i < 3$ discarded. These interval widths and coverage compare favorably to those of the full hierarchical model, listed in Appendix C, with the average Morris corrected empirical Bayesian interval width only slightly larger than the full hierarchical average interval width.

3.4 Conclusion

Some results have been shown in extending the interval comparison framework to the full NEFQVF family. Simulations from the gamma-Poisson model mirrors the results from the beta-binomial model. In particular, the concept of the Morris hyperprior is introduced that allows for the extension of the conjugate prior to the hierarchical setting in a limited fashion. This conjugacy is used to derive a new method of correcting empirical Bayesian intervals, which shows promise in terms of interval width and coverage.

Though approximations are sufficient for the purposes of this article, it is clear from simulations that approximate conjugate distribution of equation (3.9) is strong. Furthermore, though the shrinkage parameter B assumed known, it may be possible to extend the results to scenarios with unknown B — Morris (1988) uses Edgeworth expansions to estimate a posterior distribution that accounts for an unknown B, and shows that in the normal-normal case this is the corrected intervals of Morris (1983b). How this performs in the non-normal scenario is unknown, however, and it is furthermore unkown if there exists a hyperprior on M (or some function of M) that may recreate the hierarchical conjugacy of the Morris hyperprior in the scenario with known B. It is clear there exists a great deal of theory in the area of hierarchical modeling in the natural exponential family with quadratic variance function that is waiting to be discovered.

CHAPTER 4. BAYESIAN AND EMPIRICAL BAYESIAN ESTIMATION OF A BASEBALL TEAM'S WINNING PERCENTAGE USING THE ZERO-INFLATED GEOMETRIC DISTRIBUTION

4.1 Winning Percentage Estimators

Developed in Bill James's abstract in James (1983), the pythagorean expectation is commonly used to estimate the "true" talent level of a baseball team. Supposing that the number of games won by a team follows some random process with a fixed expectation, the pythagorean expectation attempts to estimate the expected winning proportion given the total number of runs scored and allowed. A team's "true" winning percentage p is estimated as

$$p = \frac{\text{Runs Scored}^2}{\text{Runs Scored}^2 + \text{Runs Allowed}^2} \tag{4.1}$$

In Miller (2007), this formula is derived by placing independent Weibull distributions on runs scored and allowed per game. More modern versions of this estimator follow the same form but use an estimated exponent. The "pythagenpat" estimator developed by sabermetrician David Smyth and pseudonymous sabermetrician Patriot determines the exponent x by

$$x = \left(\frac{\text{Runs Scored} + \text{Runs Allowed}}{\text{Number of Games Played}}\right)^{0.287} \tag{4.2}$$

and estimates a team's true winning percentage as

$$p = \frac{\text{Runs Scored}^x}{\text{Runs Scored}^x + \text{Runs Allowed}^x} \tag{4.3}$$

This has empirically shown to be an improvement over the exponent of 2 used in James' original estimator in terms of mean squared error from the actual winning percentage, as seen in Table 4.4.

Many other estimators exist in both the public and private spheres that modify these forms, for example by estimating the expected number of runs a team should have scored and allowed given the counts of plate-appearance events such as singles, doubles, walks, and home runs.

4.2 Distribution of Runs Scored and Runs Allowed

These estimators, however, have traditionally relied upon the total of runs scored and runs allowed with lesser consideration for the distribution of runs scored and runs allowed per inning. Efforts to determine per-inning distributions for runs scored and allowed have proceeded slowly. In Tango (2001), pseudonymous sabermetrician Tom Tango used a zero-inflated geometric distribution (which he refers to as a "Tango" distribution) to describe the distribution of runs scored and allowed per inning, using a moment-based estimation procedure which assumes a known average runs per game. This method was further described and expanded upon in Glass and Lowry (2008), where it is referred to as a "pseudogeometric" distribution. Albert (2015) describes statistical efforts further to determine run-scoring distributions per inning, particularly those involving regression-based techniques, but fails to include the work of Tango, Glass, and Lowry, potentially because run scoring per inning distributions remain relatively unused within the sabermetric community and because of the differing nomenclature from the statistical community.

Fitting the zero-inflated geometric distributions with a maximum likelihood estimation procedure allows for a more accurate fit of distributions to empirical data. Begin by defining RS_i and RA_i to be the number of runs scored and allowed in inning i, respectively. Each is independently assumed to follow a zero-inflated geometric distribution with density

$$p(k) = \begin{cases} \phi + (1-\phi)p & k = 0 \\ (1-\phi)(1-p)^k p & k > 0 \end{cases} \tag{4.4}$$

Assuming an independent sample from this distribution, define n as the total number of innings observed and n_0 as the number of innings observed with $k = 0$ runs scored or allowed, depending on which distribution is being fit. The log-likelihood function is given by

$$\ell(p,\phi) = n_0 \log(\phi + (1-\phi)p) + (n-n_0)\log(1-\phi) + \left(\sum_{i=1}^{n} k_i\right)\log(1-p) + (n-n_0)\log(p) \quad (4.5)$$

Equation (4.5) may be maximized by standard optimization techniques, which should be well-behaved. The choice of $p = 0.5$ and $\phi = 0.5$ as starting values for a Newton-Raphson algorithm in particular has not presented any computational difficulties.

The estimation procedure will result in two independent two-parameter distributions.

$$RS_i \overset{iid}{\sim} ZIGeo(\hat{p}_{RS}, \hat{\phi}_{RS}) \quad (4.6)$$

$$RA_i \overset{iid}{\sim} ZIGeo(\hat{p}_{RA}, \hat{\phi}_{RA}) \quad (4.7)$$

4.3 Winning Percentage Estimator

Rather than relying on the total runs scored and allowed, a superior winning percentage estimator may be constructed directly from the fitted distributions in equations (4.6) and (4.7). The winning percentage is determined by forcing the runs scored per inning distribution to "play" the runs allowed per inning distribution in a simulated game:

1. Simulate 9 innings each from the zero-inflated geometric distributions fit to a given team's runs scored per inning and runs allowed per inning distributions.

2. If the simulated number of runs scored in nine inning is larger than the simulated number of runs allowed in nine innings, declare the game to be a "Win." If the number of runs scored is smaller than the number of runs allowed, declare the game to be a "Loss." If the sum number of runs scored and allowed are equal, proceed to the next step.

3. Simulate one inning from each of the zero-inflated geometric distributions fit to a given team's runs scored per inning and runs allowed per inning distributions.

4. If the simulated number of runs scored in the extra inning is larger than the simulated number of runs allowed in the extra inning, declare the game to be a "Win." If the number of runs scored is smaller than the number of runs allowed in the extra inning,

declare the game to be a “Loss.” If the number of runs scored and allowed in the extra inning is equal, return to the previous step.

This process produces a $P(Win)$ with formula given by

$$P(Win) = P\left(\sum_{i=1}^{9} RS_i > \sum_{i=1}^{9} RA_i\right) + P\left(\sum_{i=1}^{9} RS_i = \sum_{i=1}^{9} RA_i\right)\left[\frac{P(RS_i > RA_i)}{1 - P(RS_i = RA_i)}\right] \tag{4.8}$$

The derivation of this formula is shown in Appendix E, and was also given in Bukiet et al. (1997) in the context of Markov chain analysis of baseball games. Following the development of Glass and Lowry (2008) (and partial development of Ben Vollmayr-Lee in Tango (2001)), the distribution of runs per nine innings has mass function

$$p(k) = \begin{cases} [\phi + (1-\phi)p]^9 & k = 0 \\ \sum_{i=1}^{\min(9,k)} \binom{9}{i}\binom{k-1}{i-1}[\phi + (1-\phi)p]^{9-i}(1-\phi)^{i-1}p^i(1-p)^k & k > 0 \end{cases} \tag{4.9}$$

The quantities needed for the estimator can then be calculated as

$$P(RS_i = RA_i) = \sum_{k=0}^{\infty} p_{RS}(k)p_{RA}(k) \tag{4.10}$$

$$P(RS_i > RA_i) = \sum_{k=1}^{\infty} p_{RS}(k)\left[\sum_{j=0}^{k} p_{RA}(j)\right] \tag{4.11}$$

Where $p_{RS}(k)$ and $p_{RA}(k)$ are the probability mass functions of the fit zero-inflated geometric densities in equations (4.6) and (4.7), respectively. These formulas may also be applied using the mass function of the nine-inning sum given by equation (4.9) in order to obtain the probability that the game is tied or won after nine innings. Though the sums are infinite, summing from $k = 0$ to a finite number is quick and accurate as a numerical approximation — tests suggest $k = 20$ works well.

4.4 Model Fit

As previously noted, fitting zero-inflated geometric distributions to the distributions of runs scored and allowed per inning by maximum likelihood presents a very close approximation to the empirical distributions. For the 2015 Atlanta Braves of the National league, the resulting models are

$$RS_i \sim ZIGeo(0.585, 0.441) \tag{4.12}$$

$$RA_i \sim ZIGeo(0.557, 0.334) \tag{4.13}$$

Empirical runs per inning and modeled runs per inning appear, at least to the naked eye, very close.

Table 4.1 Empirical and Zero-Inflated Geometric Distributions of Runs Scored per Inning for the 2015 Atlanta Braves

k	0	1	2	3	4	5	6	7
Empirical $p(k)$	0.768	0.132	0.062	0.021	0.010	0.004	0.000	0.002
Model $p(k)$	0.768	0.136	0.056	0.023	0.010	0.004	0.002	0.001

The model fits equally well to the empirical distribution of runs allowed per inning.

Table 4.2 Empirical and Zero-Inflated Geometric Distributions of Runs Allowed per Inning for the 2015 Atlanta Braves

k	0	1	2	3	4	5	6	7
Empirical $p(k)$	0.705	0.163	0.071	0.036	0.016	0.006	0.002	0.001
Model $p(k)$	0.705	0.164	0.073	0.032	0.014	0.006	0.003	0.001

The maximum likelihood method of estimation perfectly matches the probability of zero runs scored for each distribution. Furthermore, the expected number of runs scored and allowed per game for the fitted distribtions is exactly equal to the sample average number of runs scored and allowed per inning.

4.5 Comparisons

The pythagorean expectation and pythagenpat winning proportion estimators given by equations (4.1) and (4.3) were calculated for each team in the 2015 major league baseball season. Zero-inflated geometric distributions were also fit to the distributions of runs scored and allowed for each team, and the zero-inflated geometric estimate given by equation (4.8) was calculated. The results are shown below in Table 4.3, converted to a 162 game scale.

Table 4.3 Observed Wins, Pythagorean Expectation (P.rean), Pythagenpat Expcectation (P.pat), and Zero-Inflated Win Probability (ZI) for the 2015 MLB Season

Team	Wins	P.rean	P.pat	ZI	Team	Wins	P.rean	P.pat	ZI
ARI	79	81.79	81.74	82.42	MIL	68	71.49	72.22	72.64
ATL	67	58.71	60.59	60.90	MIN	83	80.54	80.57	80.09
BAL	81	83.30	83.13	82.57	NYM	90	89.72	88.90	89.15
BOS	78	80.46	80.49	80.91	NYY	87	88.30	87.83	86.39
CHC	97	91.08	90.12	90.07	OAK	68	77.02	77.30	77.20
CHW	76	71.36	72.22	73.08	PHI	63	60.67	62.02	61.95
CIN	64	67.84	68.83	68.96	PIT	98	93.58	92.38	91.57
CLE	81	84.07	83.74	84.36	SDP	74	71.53	72.27	71.96
COL	68	70.09	70.55	69.62	SEA	76	72.82	73.46	74.06
DET	74	68.27	68.96	70.20	SFG	84	89.43	88.67	87.80
HOU	86	94.26	93.14	92.78	STL	100	97.68	95.71	94.99
KCR	95	90.81	90.01	90.08	TBR	80	81.25	81.23	81.87
LAA	85	79.30	79.45	79.70	TEX	88	82.96	82.85	83.31
LAD	92	90.21	89.28	88.84	TOR	93	103.48	102.49	101.12
MIA	71	72.86	73.65	74.53	WSN	83	89.21	88.50	88.10

The zero-inflated geometric estimator compares positively to the Pythagorean expectation of James (1983) and the pythagenpat estimator.

To compare by more than just the naked eye, define the root mean-squared error from the actual winning percentage as

$$RMSE = \sqrt{\sum_{t=1}^{30}(N_t p_t - W_t)^2} \tag{4.14}$$

where t indexes team from one to thirty, N_t is the number of games that the team has played (which is 162 for every team in the 2015 season, but may be different, as there is the potential for teams out of playoff contention to choose not to play previously postponed games, and rarely there are ties), p_t is the estimated winning percentage using some expectation method, and W_t is the actual number of wins. The root mean-squared error for 2015 over all 30 teams is 4.77 games for the pythagorean expectation (using the default exponent of two), compared to 4.65 for the pythagenpat expectation and 4.52 for the zero-inflated geometric expectation.

In order to determine if this smaller root mean squared error for the zero-inflated geometric estimator holds in multiple years, the number of runs scored and allowed in each inning for all 30 teams was collected for each major league baseball season from 2000 to 2015. Again, standard pythagorean expectation (using the default exponent of two), pythagenpat expectation, and zero-inflated geometric expectation (each converted to wins) were compared. The mean-squared error over all teams for each method in each year is shown below in Table 4.4, as well as the average mean-squared error and root average mean-squared error over all sixteen seasons. Noticeably, the zero-inflated geometric expectation has the smallest mean-squared error in all years except for 2004, 2006, and 2012, all of which had the standard pythagorean expectation as the smallest mean-squared error.

Table 4.4 Mean squared errors for all baseball teams for the pythagorean, pythagenpat, and zero-inflated geometric winning percentage estimation techniques from 2000 to 2015

Year	Pythagorean	Pythagenpat	ZI
2000	10.55	10.65	8.71
2001	15.43	15.32	12.71
2002	16.87	16.98	15.70
2003	14.67	15.23	13.22
2004	17.38	18.39	19.23
2005	21.07	19.79	15.26
2006	15.65	15.91	15.75
2007	18.63	17.43	12.97
2008	17.65	17.52	17.19
2009	22.55	21.59	19.02
2010	11.16	7.69	6.78
2011	15.61	15.25	12.24
2012	13.71	14.15	14.14
2013	15.11	13.27	10.61
2014	15.63	14.42	12.24
2015	22.78	21.58	20.44
MSE	16.53	15.95	14.14
RMSE	4.07	3.99	3.76

Comparing average mean-squared error and root average mean-squared error, the zero-inflated geometric estimator outperforms the pythagenpat estimator by more than the pythagenpat estimator outperforms the pythagorean expectation.

4.6 Interval Estimation

It is neither immediately clear nor simple how to obtain interval estimates for winning proportion estimators based upon total runs scored and allowed. Tung (ND) uses a parametric bootstrap to obtain confidence intervals for the pythagorean expectation using the parametric model of Miller (2007), as well as investigating logistic regression techniques. Intervals for other estimators such as pythagenpat have not been developed. However, Interval estimation for the zero-inflated geometric estimator is fairly straightforward .

Using a Newton-Raphson estimation technique for maximum likelihood estimation, covariance matrices between $\hat{p}$ and $\hat{\phi}$ should be readily available. As RS_i and RA_i are both being fit to independent zero-inflated geometrics, two 2x2 covariance matrices V_{RS} and V_{RA} will be obtained. Define the full covariance matrix V as

$$V = \begin{bmatrix} V_{RS} & 0_{2x2} \\ 0_{2x2} & V_{RA} \end{bmatrix} \tag{4.15}$$

Furthermore, define the estimated winning percentage as a function of the estimated parameters:

$$h(\hat{p}_{RS}, \hat{\phi}_{RS}, \hat{p}_{RA}, \hat{\phi}_{RA}) = P(Win) \tag{4.16}$$

where $P(Win)$ is as in equation (4.8). Using the delta method, a variance estimate for the zero-inflated geometric win proportion estimator is provided by

$$(\nabla h)^T V (\nabla h) \tag{4.17}$$

As h is a complex function that can not be differentiated by hand, numerical methods must be used to estimate the gradient ∇h. The resulting variance in equation (4.17) is for the zero-inflated geometric winning proportion - to convert to the observed number of games over the course of a 162 game season, multiply the result by 162^2.

Interval estimation in a Bayesian framework is also straightforward. Placing priors upon $p_{RS}, \phi_{RS}, p_{RA}$, and ϕ_{RA}, standard Markov chain Monte Carlo (MCMC) techniques will give a sample from the posterior distribution of each parameter — Carlin and Louis (2000) provides a reference for such techniques. These samples may be inserted directly into equation (4.16) in order to produce a posterior predictive density for winning percentage, from which central or highest posterior density intervals may be constructed.

To illustrate, the following priors were placed on the parameters for the zero-inflated geometric distributions of runs scored and allowed for all 30 major league baseball teams of the 2015 season:

$$p_{RS}, p_{RA} \sim N(0.552, 0.020^2) \tag{4.18}$$

$$\phi_{RS}, \phi_{RA} \sim N(0.380, 0.033^2) \tag{4.19}$$

$$\tag{4.20}$$

This combination was chosen to produce a prior predictive distribution of winning probability that is approximately normally distributed with mean 0.5 and standard deviation 0.057, corresponding to an approximately 95% probability of winning between 62 and 100 games over the course of a 162 game major league baseball season.

Intervals (using a nominal coverage of 95%) using both the maximum likelihood and Bayesian estimation techniques for the zero-inflated geometric winning percentage (converted to a 162 game scale) are given in Table 4.5 below. For the MCMC, A single chain of 2500000 draws was obtained for each of the distributions of runs scored and allowed for each team after smaller-scale testing failed to indicate any problems with convergence and showed excellent mixing. Starting values of $p = 0.5$ and $\phi = 0.5$ were used for each MCMC chain. To obtain Bayesian intervals, standard 2.5% and 97.5% quantiles were taken from the posterior predictive distributions of winning percentage for each team.

Table 4.5 95% Maximum Likelihood and Central Bayesian Intervals for the Winning Total of MLB Teams in the 2015 Season

Team	ML CI	Bayes CI	Team	ML CI	Bayes CI
ARI	(72.20, 92.63)	(72.58, 91.65)	MIL	(62.42, 82.87)	(65.29, 84.47)
ATL	(51.03, 70.77)	(56.35, 75.28)	MIN	(69.83, 90.36)	(70.62, 89.61)
BAL	(72.26, 92.88)	(72.45, 91.60)	NYM	(78.95, 99.34)	(77.69, 96.60)
BOS	(70.68, 91.14)	(71.35, 90.31)	NYY	(76.20, 96.58)	(75.28, 94.26)
CHC	(79.88, 100.26)	(78.09, 97.20)	OAK	(66.99, 87.42)	(68.49, 87.78)
CHW	(62.89, 83.26)	(65.73, 84.82)	PHI	(55.03, 74.70)	(57.17, 75.69)
CIN	(58.89, 79.03)	(62.45, 81.55)	PIT	(81.49, 101.65)	(79.17, 98.31)
CLE	(74.07, 94.65)	(74.48, 93.55)	SDP	(61.75, 82.17)	(64.25, 83.63)
COL	(59.49, 79.76)	(62.76, 81.64)	SEA	(63.89, 84.23)	(66.39, 85.10)
DET	(60.17, 80.23)	(63.68, 82.56)	SFG	(77.54, 98.06)	(76.32, 95.55)
HOU	(82.61, 102.94)	(80.36, 99.22)	STL	(84.89, 105.09)	(81.88, 100.91)
KCR	(79.88, 100.27)	(78.37, 97.21)	TBR	(71.57, 92.17)	(72.22, 91.30)
LAA	(69.38, 90.02)	(70.31, 89.44)	TEX	(73.05, 93.57)	(73.40, 92.33)
LAD	(78.58, 99.11)	(77.31, 96.37)	TOR	(91.35, 110.89)	(86.92, 105.35)
MIA	(64.22, 84.84)	(66.67, 85.76)	WSN	(77.82, 98.38)	(76.92, 95.95)

The moderately informative priors given by equations (4.18) and (4.20) have a tendency to shrink intervals for teams that performed either extremely well or extremely poor towards a mean of 81 games. For example, the Toronto Blue Jays and Atlanta Braves (zero-inflated geometric winning total estimates of 101.12 and 60.59 games, respectively) saw their intervals moved by approximately 5 games towards the central 81 game mark. The Tampa Bay Rays, however (zero-inflated geometric winning total estimate of 81.87 games) saw only an approximate 1 game change in their interval.

The Bayesian method can also be used for midseason prediction by fitting the distributions to the observed data up to a certain number of games and using draws from the posterior predictive distribution for winning percentage in equation (4.16) to simulate the number of wins in the remaining games from a basic binomial distribution. Using data from 2000 to 2015, this predictive method was tested using the priors in equations (4.18) and (4.20) at 40, 81, and 120 games, and taking 95% quantiles from the predictive interval for win total (adding the current number of wins to the simulated number of wins over the remaining games) produced empirical coverage of 93.75%, 92.08%, and 93.75%, respectively.

4.7 An Empirical Bayesian Approach to Estimation

The Bayesian approach in Section uses common priors for the parameters of each team in each year. These priors may be chosen subjectively or through an ad-hoc analysis of several years of previous $\hat{p}$ and $\hat{\phi}$ values over several years, as was the case for the priors chosen in equations (4.18) and (4.20). It may be more reasonable, however, to assume that there exists a distribution of p and ϕ values that is different for each year. A full hierarchical Bayesian analysis would choose priors for the distributions of p and ϕ values fit the model using standard MCMC methods. An alternative is to estimate the prior values themselves from the data — this is an empirical Bayesian technique.

Begin by defining zero-inflated geometric distributions for runs scored per inning and runs allowed per inning.

$$RS_i \overset{iid}{\sim} ZIGeo(p_{RS}, \phi_{RS}) \tag{4.21}$$

$$RA_i \overset{iid}{\sim} ZIGeo(p_{RA}, \phi_{RA}) \tag{4.22}$$

For priors, normal distributions are suggested, both for computational ease and because histograms of the maximum likelihood estimates of p and ϕ for all 30 teams shows visual evidence of being unimodal and bell-shaped, especially when considering multiple years worth of estimates. A common prior may be used for both p_{RS} and p_{RA} and for both ϕ_{RS} and ϕ_{RA}.

$$p_{RS}, p_{RA} \sim N(\mu_p, \sigma_p^2) \tag{4.23}$$

$$\phi_{RS}, \phi_{RA} \sim N(\mu_\phi, \sigma_\phi^2) \tag{4.24}$$

This choice of using common priors for both p values and both ϕ values follows from two sources - first, constructing separate histograms of maximum likelihood estimates of p for runs scored and allowed shows an incredibly strong similarity between the two distributions, and the similarity is also seen with separate histograms of the maximum likelihood estimates of ϕ for runs scored and allowed (especially when considering multiple years worth of estimates). Second, reasoning from knowledge of baseball that each run scored for a team corresponds to a run allowed for another team suggests that there should be a strong, if not equal, mathematical connection between the two. The choice of the normal distribution follows from the symmetric and unimodal shape of histograms of the maximum likelihood estimates of p and ϕ for runs scored and allowed.

Direct maximization of the full likelihood function is not numerically tractable. An alternative is to use the EM algorithm to iteratively solve for the parameters values. A brief description of the method for this specific model will be given, but full details of the EM algorithm in empirical Bayesian estimation may be found in Carlin and Louis (2000).

Starting parameter estimates $\mu_p^{(0)}, \sigma_p^{(0)}, \mu_\phi^{(0)}$, and $\sigma_\phi^{(0)}$ are required - one choice is to take the sample mean and sample standard deviation of the maximum likelihood estimates $\hat{p}$ and $\hat{\phi}$ (both for runs scored and allowed) over all teams. Sufficient statistics for μ and σ of a normal distribution are given by the first two sample moments. For the prior distribution $N(\mu_p, \sigma_p^2)$, these are

$$\frac{1}{60}\left(\sum_{t=1}^{30} p_{RS,t} + \sum_{t=1}^{30} p_{RA,t}\right) \tag{4.25}$$

$$\frac{1}{60}\left(\sum_{t=1}^{30} p_{RS,t}^2 + \sum_{t=1}^{30} p_{RA,t}^2\right) \tag{4.26}$$

where t indexes team. Since p_{RS} and p_{RA} share a common prior, both may be considered observations from the same distribution, and hence the sufficient statistics will include both values - sixty in total. Similarly, the first two sample moments of the combined set of ϕ_{RS} and ϕ_{RA} are sufficient for μ_ϕ and σ_ϕ.

The E-step may be found by by calculating the expected values of the statistics in equations (4.25) and (4.26) from each of the posteriors $p(p_{RS}, \phi_{RS}|\mu_p^{(j)}, \sigma_p^{(j)}, \mu_\phi^{(j)}, \sigma_\phi^{(j)})$ and $p(p_{RA}, \phi_{RA}|\mu_p^{(j)}, \sigma_p^{(j)}, \mu_\phi^{(j)}, \sigma_\phi^{(j)})$. These expectations must be calculated for all 30 teams. The sufficient statistics, $E[p]$ and $E[p^2]$ must be calculated separately — $E[p]^2$ should not be used to estimate $E[p^2]$, and similarly for $E[\phi]$ and $E[\phi^2]$.

For the M-step, calculate new prior parameter estimates from the sufficient statistics.

$$\mu_p^{(j+1)} = \frac{1}{60}\left(\sum_{t=1}^{30} p_{RS,t} + \sum_{t=1}^{30} p_{RA,t}\right) \tag{4.27}$$

$$\sigma_p^{(j+1)} = \sqrt{\frac{1}{60}\left(\sum_{t=1}^{30} p_{RS,t}^2 + \sum_{t=1}^{30} p_{RA,t}^2 - \frac{1}{60}\left(\sum_{t=1}^{30} p_{RS,t} + \sum_{t=1}^{30} p_{RA,t}\right)^2\right)} \tag{4.28}$$

The value t again indexes team. New estimates $\mu_\phi^{(j+1)}$ and $\sigma_\phi^{(j+1)}$ may similarly be acquired by using ϕ and ϕ^2 posterior expectations in equations (4.27) and (4.28) above.

This EM algorithm was applied to data from the 2015 MLB season. For the E-step, posteriors for all 30 teams were calculated via MCMC, though this step may be made more efficient through more specific sampling techniques. A total of 100000 draws from each MCMC chain was used at each E-step. Stopping criterion must necessarily be looser because of the effect of MCMC variation in the E-step, so iteration was halted when the first three nonzero digits of each prior parameter estimate failed to change in multiple successive iterations. Estimates of the prior distributions are given by

$$p_{RS}, p_{RA} \sim N(0.5609, 0.0154^2) \tag{4.29}$$

$$\phi_{RS}, \phi_{RA} \sim N(0.3953, 0.0238^2) \tag{4.30}$$

These may then be used to produce interval estimates for the team winning percentage through standard Bayesian calculations as described in Section 4.7. As an example, in 2015 the Toronto Blue Jays had a 95% empirical Bayesian posterior interval for winning total between

(85.83, 102.76), the Atlanta Braves had an empirical Bayesian interval for winning total of (59.63, 76.73), and the Tampa Bay Rays had an empirical Bayesian interval for winning total of (72.83, 90.23). The effect of the empirical Bayesian analysis for this particular year was to pull the intervals even more towards the center of 81 games than the Bayesian intervals given in Table 4.5.

4.8 Conclusion and Further Work

The use of parametric modeling of the distributions of runs scored and allowed by the zero-inflated geometric distribution presents interesting new avenues for research. One such avenue is presented here - an estimator for winning percentage may be constructed that is superior in terms of mean squared error to estimators that rely on total runs scored and allowed, such as the Pythagenpat estimator. Furthermore, the parametric nature of the model allows for relatively direct interval estimation through standard frequentist, Bayesian, or empirical Bayesian theory, rather than relying on purely monte carlo methods such as bootstrapping.

This model makes several strong assumptions, at least one of which is known to be violated. Empirical evidence suggests that the independence of the zero-inflated geometric distributions between runs scored and allowed is appropriate; however, assuming independence between innings is clearly inappropriate - pitchers typically throw for multiple innings, so an elite pitcher that allows no runs in one inning is more likely to prevent runs in future innings, and conversely, a poor pitcher that allows multiple runs in one inning is more likely to allow multiple runs in additional innings (or possibly be removed from the game entirely). Models that account for the dependence between innings, or allow for effects that are known to influence run scoring distributions - such as playing at home or away - could potentially be fit and offer further improvements and shorter interval widths.

BIBLIOGRAPHY

Albert, J. (2015). Beyond Runs Expectancy. *Journal of Sports Analytics*, 1(1):3–18.

Bukiet, B., Harold, E., and Palacios, J. (1997). A Markov Chain Approach to Baseball. *Operations Research*, 45(1):14 – 23.

Carlin, B. and Gelfand, A. (1990). Approaches for Empirical Bayes Confidence Intervals. *Journal of the American Statistical Association*, 68(409):105–114.

Carlin, B. and Gelfand, A. (1991). A Sample Reuse Method for Accurate Parametric Empirical Bayes Confidence Intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(1):189 – 200.

Carlin, B. and Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall \CRC.

Casella, G. and Berger, R. (2002). *Statistical Inference, Second Edition.* Duxbury Press.

Deely, J. and Lindley, D. (1991). Bayes Empirical Bayes. *Journal of the American Statistical Association*, 76(76):833–841.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press.

Efron, B. and Morris, C. (1973). Stein's Estimation Rule and Its Competitors – An Empirical Bayes Approach. *Journal of the American Statistical Association*, 68(341):117–130.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association*, 70(350):311–319.

Ghosh, M. and Liu, R. (2011). Moment Matching Priors. *Sankhyā: The Indian Journal of Statistics, Series A*, 73:185–201.

Glass, D. and Lowry, J. (2008). Quasigeometric Distributions and Extra Inning Baseball Games. *Mathematics Magazine*, 81:127–137.

James, B. (1983). Bill James' Baseball Abstract 1983. Ballantine.

Kass, R. and Steffey, D. (1989). Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models). *Journal of the American Statistical Association*, 84(407):717–726.

Kleinman, J. (1973). Proportions with Extraneous Variance: Single and Independent Sample. *Journal of the American Statistical Association*, 68(341):46 – 54.

Laird, N. and Louis, T. (1987). Empirical Bayes Confidence Intervals Based on Bootstrap Samples. *Journal of the American Statistical Association*, 82(399):739–750.

Lee, J. and Sabavala, D. (1987). Bayesian Estimation and Prediction for the Beta-Binomial Model. *Journal of Business & Economic Statistics*, 5(3):357 – 367.

Miller, S. (2007). A Derivation of the Pythagorean Won-Loss Formula in Baseball. *Chance Magazine*, 20(7):40 – 48.

Morris, C. (1982). Natural Exponential Families with Quadratic Variance Functions. *The Annals of Statistics*, 10(1):65–80.

Morris, C. (1983a). Natural Exponential Families with Quadratic Variance Functions: Statistical Theory. *The Annals of Statistics*, 11(2):515–529.

Morris, C. (1983b). Parametric Empirical Bayes Confidence Intervals. In *Scientific Inference, Data Analysis, and Robustness*, pages 25–50.

Morris, C. (1983c). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, 78(381):47 – 55.

Morris, C. (1988). Determining the Accuracy of Bayesian Empirical Bayes Estimates in the Familiar Exponential Families. In *Statistical Decision Theory and Related Topics IV*, volume 1, pages 251 – 263.

Morris, C. and Lock, K. (2009). Unifying the Named Natural Exponential Families and their Relatives. *The American Statistician*, 63(3):248–253.

Robbins, H. (1956). An Empirical Bayes Approach to Statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–163.

Robbins, H. (1964). The Empirical Bayes Approach to Statistical Decision Problems. *The Annals of Mathematical Statistics*, 35(1):1–20.

Stein, J. (1962). Confidence Sets for the Mean of a Multivariatie Normal Distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):265–296.

Tango, T. (2001). http://www.tangotiger.net/files/tangodist.zip.

Tung, D. (N.D.). Confidence Intervals for the Pythagorean Formula in Baseball.

Walter, G. and Hamedani, G. (1987). Empiric Bayes Estimation of a Binomial Probability. *Communications in Statistics – Theory and Methods*, 16(2):559 – 577.

Walter, G. and Hamedani, G. (1991). Bayes Empirical Bayes Estimation for Natural Exponential Families with Quadratic Variance Functions. *The Annals of Statistics*, 19(3):1191 – 1224.

Young-Xu, Y. and Chan, K. A. (2008). Pooling Overdispersed Binomial Data to Estimate Event Rate. *BMC Medical Research Methodology*, 8:58.

APPENDIX A. ESTIMATION PROCEDURES

The marginal density of μ and M for the model given in equations (2.7) is given below in equation (A.1).

$$p(y_i|\mu, M) = \int_0^1 \binom{n_i}{y_i} \frac{\theta_i^{y_i+\mu M-1}(1-\theta_i)^{n_i-y_i+(1-\mu)M-1}}{\beta(\mu M, (1-\mu)M)} d\theta_i = \binom{n_i}{y_i} \frac{\beta(y_i+\mu M, n_i-y_i+(1-\mu)M)}{\beta(\mu M, (1-\mu)M)} \tag{A.1}$$

And for a sample of size k observations, the marginal maximum likelihood estimate is given by maximizing the log-likelihood function

$$\ell(\mu, M) = \left[\sum_{i=1}^{k} \log(\beta(x_i+\mu M, n_i-x_i+(1-\mu)M))\right] - k\log(\beta(\mu M, (1-\mu)M))$$

For the method of moments estimator, the moments of the two-stage model can then be written as

$$E\left[E\left[\frac{y_i}{n_i}\middle|\theta_i\right]\right] = E[\theta_i] = \mu$$
$$Var\left(\frac{y_i}{n_i}\right) = E\left[Var\left(\frac{y_i}{n_i}\middle|\theta_i\right)\right] + Var\left(E\left[\frac{y_i}{n_i}\middle|\theta_i\right]\right) = \frac{\mu(1-\mu)}{n_i}\left[1+\frac{n_i-1}{M+1}\right]$$

Assuming there are k observations, setting the sample mean and variance of the y_i/n_i equal to the theoretical mean and variance and solving for μ and M gives

$$\hat{\mu} = \frac{\sum y_i}{\sum n_i} \tag{A.2}$$

And defining $\hat{\theta}_i = y_i/n_i$, the estimator for $\hat{M}$ is given as

$$\hat{M} = \frac{\hat{\mu}(1-\hat{\mu}) - s^2}{s^2 - \frac{\hat{\mu}(1-\hat{\mu})}{k}\sum(1/n_i)}$$

where

$$s^2 = \frac{k\sum n_i(\hat{\theta}_i - \hat{\mu})^2}{(k-1)\sum n_i}$$

This is a very naive estimator. An alternative moments-based estimator is proposed in Kleinman (1973) that iteratively weights the observations until convergence. Start by defining

$$\hat{\theta}_i = \frac{y_i}{n_i}$$

and estimate the mean as

$$\hat{\mu} = \frac{\sum w_i\hat{\theta}_i}{w}$$

where w_i are weights and $w = \sum w_i$. Initial values of weights must be chosen - using $w_i = n_i$ is a common choice, and in the initial step will give the naive estimator in equation (A.2). Then defining $S = \sum w_i(\hat{\theta}_i - \hat{\mu})^2$, The dispersion parameter ϕ is estimated as

$$\hat{\phi} = \frac{S - \hat{\mu}(1-\hat{\mu})\left[\sum \frac{w_i}{n_i}\left(1 - \frac{w_i}{w}\right)\right]}{\hat{\mu}(1-\hat{\mu})\left[\sum w_i\left(1 - \frac{w_i}{w}\right) - \sum \frac{w_i}{n_i}\left(1 - \frac{w_i}{w}\right)\right]}$$

The weights w_i are then re-estimated as

$$w_i = \frac{n_i}{1 + \hat{\phi}(n_i - 1)}$$

The process is iterated until the different in weights between iterations is smaller than a specified tolerance. When all sample sizes are equal, the naive moments estimator produces the same estimates as the iterated moments estimator, and can simply be referred to as the method of moments estimate. When sample sizes differ, the iterated moments estimator produces estimates that tend to be closer in value to the estimates produced by maximum likelihood.

APPENDIX B. TERBINAFINE DATA AND ANALYSIS

Table B.1 Terbinafine Trials Data and Raw Adverse Reaction Proportion

Arm	n_i	y_i	$\hat{\theta}_i$	Arm	n_i	y_i	$\hat{\theta}_i$	Arm	n_i	y_i	$\hat{\theta}_i$
1	184	7	3.80%	21	142	2	1.41%	41	120	3	2.50%
2	65	1	1.54%	22	124	8	6.45%	—	—	—	—
3	33	1	3.03%	23	56	1	1.79%	—	—	—	—
4	151	4	2.65%	24	12	0	0.00%	—	—	—	—
5	24	0	0.00%	25	50	0	0.00%	—	—	—	—
6	30	0	0.00%	26	88	3	3.41%	—	—	—	—
7	20	0	0.00%	27	48	0	0.00%	—	—	—	—
8	22	0	0.00%	28	75	4	5.33%	—	—	—	—
9	50	4	8.00%	29	76	0	0.00%	—	—	—	—
10	50	5	10.00%	30	56	1	1.79%	—	—	—	—
11	18	0	0.00%	31	153	9	5.88%	—	—	—	—
12	26	0	0.00%	32	68	1	1.47%	—	—	—	—
13	72	0	0.00%	33	120	13	10.83%	—	—	—	—
14	30	1	3.33%	34	44	0	0.00%	—	—	—	—
15	16	0	0.00%	35	84	0	0.00%	—	—	—	—
16	26	2	7.69%	36	21	0	0.00%	—	—	—	—
17	95	8	8.42%	37	145	3	2.07%	—	—	—	—
18	95	3	3.16%	38	83	10	12.05%	—	—	—	—
19	186	0	0.00%	39	68	3	4.41%	—	—	—	—
20	146	11	7.53%	40	30	3	10.00%	—	—	—	—

Raw counts of events and number of patients for the terbinafine trials are shown above in Table B.1. Estimates of θ_i for each method (maximum likelihood, naive method of moments, iterated method of moments, and hierarchical Bayes) are given by

Table B.2 Terbinafine Trials Empirical and Hierarchical Bayesian Estimates

Arm	$\hat{\theta}_i^{NMM}$	$\hat{\theta}_i^{IMM}$	$\hat{\theta}_i^{MLE}$	$\hat{\theta}_i^{HB}$	Arm	$\hat{\theta}_i^{NMM}$	$\hat{\theta}_i^{IMM}$	$\hat{\theta}_i^{MLE}$	$\hat{\theta}_i^{HB}$
1	0.038	0.037	0.037	0.038	21	0.023	0.019	0.018	0.018
2	0.028	0.024	0.022	0.022	22	0.053	0.056	0.058	0.058
3	0.035	0.033	0.032	0.033	23	0.03	0.026	0.024	0.024
4	0.03	0.028	0.028	0.028	24	0.033	0.028	0.026	0.025
5	0.029	0.023	0.021	0.02	25	0.024	0.017	0.014	0.014
6	0.028	0.021	0.019	0.018	26	0.036	0.034	0.034	0.034
7	0.03	0.024	0.022	0.022	27	0.024	0.017	0.015	0.014
8	0.03	0.024	0.021	0.021	28	0.044	0.046	0.047	0.048
9	0.053	0.058	0.061	0.062	29	0.02	0.013	0.011	0.011
10	0.06	0.068	0.073	0.074	30	0.03	0.026	0.024	0.024
11	0.031	0.025	0.023	0.022	31	0.051	0.053	0.054	0.054
12	0.029	0.022	0.02	0.019	32	0.027	0.023	0.022	0.021
13	0.02	0.014	0.011	0.011	33	0.078	0.088	0.091	0.092
14	0.036	0.034	0.034	0.034	34	0.025	0.018	0.016	0.015
15	0.031	0.026	0.024	0.023	35	0.019	0.012	0.01	0.01
16	0.046	0.05	0.052	0.054	36	0.03	0.024	0.022	0.021
17	0.061	0.068	0.071	0.071	37	0.027	0.024	0.023	0.023
18	0.034	0.033	0.032	0.032	38	0.077	0.09	0.094	0.096
19	0.012	0.007	0.006	0.005	39	0.04	0.04	0.041	0.041
20	0.061	0.066	0.067	0.068	40	0.053	0.06	0.064	0.067
					41	0.03	0.028	0.027	0.027

And interval widths are

Table B.3 Terbinafine Trials Empirical and Hierarchical Bayesian Interval Widths

Arm	$\hat{\theta}_i^{NMM}$	$\hat{\theta}_i^{IMM}$	$\hat{\theta}_i^{MLE}$	$\hat{\theta}_i^{HB}$	Arm	$\hat{\theta}_i^{NMM}$	$\hat{\theta}_i^{IMM}$	$\hat{\theta}_i^{MLE}$	$\hat{\theta}_i^{HB}$
1	0.045	0.048	0.05	0.05	21	0.038	0.038	0.038	0.039
2	0.051	0.054	0.055	0.056	22	0.06	0.068	0.071	0.073
3	0.064	0.076	0.08	0.084	23	0.054	0.059	0.061	0.062
4	0.043	0.046	0.046	0.047	24	0.067	0.08	0.084	0.088
5	0.06	0.066	0.068	0.069	25	0.049	0.049	0.047	0.048
6	0.057	0.061	0.061	0.062	26	0.054	0.06	0.062	0.064
7	0.063	0.07	0.072	0.074	27	0.05	0.05	0.048	0.049
8	0.061	0.068	0.07	0.071	28	0.062	0.073	0.077	0.08
9	0.073	0.091	0.099	0.105	29	0.042	0.039	0.036	0.037
10	0.078	0.099	0.108	0.116	30	0.054	0.059	0.061	0.062
11	0.064	0.072	0.075	0.077	31	0.055	0.062	0.064	0.065
12	0.059	0.065	0.065	0.067	32	0.05	0.053	0.054	0.055
13	0.043	0.04	0.038	0.039	33	0.072	0.085	0.09	0.094
14	0.066	0.078	0.084	0.088	34	0.051	0.052	0.051	0.052
15	0.065	0.075	0.078	0.081	35	0.04	0.036	0.034	0.035
16	0.075	0.097	0.107	0.117	36	0.062	0.069	0.071	0.073
17	0.069	0.082	0.087	0.09	37	0.041	0.043	0.043	0.044
18	0.052	0.057	0.059	0.06	38	0.079	0.097	0.104	0.111
19	0.025	0.02	0.018	0.019	39	0.06	0.07	0.074	0.076
20	0.061	0.069	0.072	0.074	40	0.079	0.104	0.115	0.127
					41	0.046	0.049	0.05	0.051

APPENDIX C. SIMULATIONS (KNOWN M)

Simulations were conducted in order to determine under what conditions the phenomenon of moment-based empirical Bayesian intervals being wider than hierarchical Bayesian intervals is likely to occur, to test the coverage of priors suggested under the framework, and, in general, to compare interval width correction methods based on the beta-binomial model. These simulations are in no way meant to be exhaustive, but simply an attempt to push the data-generating conditions towards different potential scenarios and observe the outcome. For each set of simulations, one hundred thousand monte carlo draws were taken for each of one thousand simulated data sets. Data sets such that $y_i = 0$ for all i or $y_i = n_i$ for all i were discarded.

Firstly, in order for the phenomenon to appear, it seems to be necessary that estimated prior mean be small - either towards one or towards zero. Simulating data from a beta-binomial model with $k = 10$ binomial observations consisting of $n_i = 10$ trials each and known $M = 4$ but data generating mean $\mu = 0.5$, coverage and average interval widths are given in Table C.1.

Table C.1 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.50, \text{Known } M = 4, n_i = k = 10$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}$	Haldane's Prior
Coverage	0.946	0.946	0.951	0.951	0.951
Width	0.452	0.452	0.46	0.46	0.46

Similarly to the baseline scenario with $\mu = 0.10$ considered in Section 2.3.6, all of the full Bayesian and empirical Bayesian analyses performed well. However, the empirical Bayesian interval using the method of moments estimator is larger than the full Bayesian analysis using

Haldane's prior only 0.24% of the time - a number which is almost certainly due simply to variance in MCMC estimation. It appears that since the phenomenon is dependent upon the shape and variance of the posterior distributions (which are dependent on the mean), moderate values of $\hat{\mu}$ which pull the sample proportion y_i/n_i towards the center of the interval $(0, 1)$ will produce intervals that are approximately symmetric and unimodal, and standard theory regarding empirical Bayesian estimation for normal distributions will do well to approximate the necessary corrections.

This can be also be seen by application of the parametric bootstrap and (unconditional) bias correction methods in Table C.2.

Table C.2 Non-Bayesian Correction Methods Coverage and Interval Width — $\mu = 0.5, \text{Known M} = 4, k = n_i = 10$

	Bootstrap MM	Bootstrap MLE	Bias Correction MM	Bias Correction MLE
Coverage	0.952	0.952	0.951	0.951
Width	0.46	0.46	0.46	0.46

All methods produce results and interval widths similar to those expected by the full hierarchical Bayesian analysis.

Furthermore, the value M must be relatively small to see the phenomenon. As $M = \alpha + \beta$ increases, the empirical Bayesian prior (and resulting posterior) will tend towards being symmetric and bell-shaped, even in cases near zero or one. Simulations with $\mu = 0.1$, $k = 10$, $n_i = 10$, and known $M = 100$ are shown in Table C.3.

Table C.3 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.10, \text{Known } M = 100, n_i = k = 10$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}$	Haldane's Prior
Coverage	0.827	0.827	0.943	0.944	0.945
Width	0.109	0.109	0.152	0.152	0.153

Not only does the empirical Bayesian interval provide relatively poor (unconditional) coverage, the interval width is generally far too short, and again the choice of prior for the Bayesian analysis does not seem to matter. In fact, in not a single instance was the empirical Bayesian interval of any kind larger than a full Bayesian analysis of any kind. The required ratio of estimated means for the phenomenon to occur is simply too large to overcome the additional variance in the hierarchical Bayesian posteriors.

The underestimation of the variance, however, may still be quantified by the ratio of empirical Bayesian interval widths using the maximum likelihood estimator to full Bayesian interval widths using Haldane's prior - the average empirical Bayesian interval width is roughly 72% that of the full Bayesian interval width. As M increases, the empirical Bayes prior and all posterior distributions will be increasingly well approximated by a normal distribution, and standard results regarding empirical Bayesian intervals based on normal theory, as discussed in Section 2.1, will apply. This is also shown in the non-Bayesian interval correction methods in Table C.4.

Table C.4 Non-Bayesian Correction Methods Coverage and Interval Width — $\mu = 0.1, \text{Known M} = 100, k = n_i = 10$

	Bootstrap MM	Bootstrap MLE	Bias Correction MM	Bias Correction MLE
Coverage	0.949	0.95	0.95	0.951
Width	0.154	0.154	0.157	0.157

Intervals are produced that are only slightly larger than the corresponding Bayesian intervals.

As the sample sizes increase, however, the phenomenon may still occur. Simulations with $\mu = 0.1$ and known $M = 4$, but with $k = 30$ observations of $n_i = 30$ trials each yield results

Table C.5 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.10, \text{Known } M = 4, n_i = k = 30$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}$	Haldane's Prior
Coverage	0.947	0.949	0.952	0.952	0.952
Width	0.152	0.152	0.152	0.152	0.152

Given that the variance parameter is known, coverage quickly converges to roughly 95% (the nominal coverage) and the empirical Bayesian intervals have an average interval width almost exactly equal to the hierarchical Bayesian interval widths, though the empirical Bayesian interval based upon the method of moments estimator is larger than the corresponding full hierarchical Bayesian interval based upon Morris-style matching of posterior MLE means using Haldane's prior roughly 38.3% of the time (though differences in interval widths are not necessarily as large and method of moments and maximum likelihood estimates will more closely agree, as compared to results of simulations using smaller sample sizes).

The larger sample size implies that most of the posteriors will be much better approximated by a normal distribution than those using a smaller sample size. Furthermore, non-Bayesian correction methods also perform well, as shown in Table C.6 below.

Table C.6 Non-Bayesian Correction Methods Coverage and Interval Width - $\mu = 0.1$, Known M $= 4, k = n_i = 30$

	Bootstrap MM	Bootstrap MLE	Bias Correction MM	Bias Correction MLE
Coverage	0.956	0.958	0.95	0.95
Width	0.152	0.152	0.154	0.153

These results combine to suggest a picture in which the phenomenon may occur for the beta-binomial moel, and is closely tied to non-Bayesian interval correction methods not producing desired results - it is necessary that $\hat{\mu}$ be either small or large, as the distributional shape of the posterior will become more dense and more skewed as it approaches the 0 and 1 boundaries. It is also necessary that M not necessarily be too large, as the empirical Bayesian interval will necessarily be much shorter than the full Bayesian interval. This implies data set with observations that tend near y_i/n_i equal to zero or one so that $\hat{M}$ is small, but concentrated near at one end, so that $\hat{\mu}$ is close to the boundaries of the beta distribution. This is a situation in which empirical Bayesian analysis is particularly useful and, as shown in the Terabifine trial data, actually has been performed.

APPENDIX D. SIMULATIONS (UNKNOWN M)

Further simulations were again conducted in order to determine the appearance of the phenomenon and to determine appropriate priors for comparison to the empirical Bayesian intervals. These are not meant to be exhaustive, but rather to investigate specific scenarios.

Intervals were constructed using quantiles from empirical Bayesian estimators using both method of moments and maximum likelihood estimators, and from intermediate hyperpriors matching the prior parameter estimates in equation (2.21) and the loosely approximate Morris-matching hyperpriors in equation (2.22). A total of one hundred thousand Monte Carlo samples were taken for each of one thousand simulated data sets for each set of generating conditions. Data sets such that $y_i = 0$ for all i or $y_i = n_i$ for all i were discarded.

Letting $\mu = 0.5$ but still with $M = 4$ and both k and n_i equal to 10 observations of 10 trials each gives results

Table D.1 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.50, M = 4, n_i = k = 10$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}, \hat{\phi}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}, \hat{\phi}_{MLE}$	Approx. Morris Matching Prior
Coverage	0.902	0.888	0.907	0.903	0.937
Width	0.427	0.418	0.438	0.435	0.454

With a true mean closer to the middle of the beta distribution support, coverage improves - the empirical Bayesian estimates now produce roughly 90% coverage (though noticeably, the method of maximum likelihood is slightly inferior to the method of moments). The set of

Haldane's prior on μ and $Beta(\mu = 0.5, M = 1)$ prior on ϕ which loosely approximates Morris-style matching works well here - in only 0.15% of cases was the maximum likelihood empirical Bayesian interval longer than the hierarchical Bayesian interval, though the method of moments empirical Bayesian intervals were longer approximately 16% of the time.

Table D.2 Non-Bayesian Correction Methods Coverage and Interval Width - $\mu = 0.5, M = 4, k = n_i = 10$

	Bootstrap MM	Bootstrap MLE	Bias Correction MM	Bias Correction MLE
Coverage	0.942	0.919	0.963	0.992
Width	0.463	0.454	0.518	0.761

Non-Bayesian correction methods do not necessarily proved an appropriate correction - the bootstrap correction to the method of moments empirical Bayesian estimator achieves roughly nominal coverage, but the correction applied to the method of maximum likelihood estimators is too short. The bias correction methods again break down and produce intervals that are far worse, in terms of width, than the hierarchical Bayesian intervals.

In the case of a smaller prior variance, with $\mu = 0.1$ and both k and n_i equal to 10 but $M = 100$, interval widths and coverage are given by

Table D.3 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.10, M = 100, n_i = k = 10$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}, \hat{\phi}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}, \hat{\phi}_{MLE}$	Approx. Morris Matching Prior
Coverage	0.87	0.868	0.951	0.952	0.976
Width	0.172	0.168	0.189	0.187	0.242

Empirical Bayesian widths are once again too short, though noticeably the intermediate hierarchical Bayesian prior which matches hyperprior moments to prior parameter estimates provides the best coverage - the approximate Morris-matching solution provides interval widths which are far too wide. Non-Bayesian methods, too, do not necessarily perform well. Maximum likelihood empirical Bayesian intervals were longer than the noninformative Bayesian intervals 0.25% of the time, and method of moments empirical Bayesian intervals were longer 0.0521% of the time. However, the empirical Bayesian intervals were longer than the intermediate hierarchical Bayesian intervals matching prior moments (which did achieve roughly nominal coverage) 40.45% and 41.20% of the time for method of moments and maximum likelihood, respectively.

Table D.4 Non-Bayesian Correction Methods Coverage and Interval Width - $\mu = 0.1, M = 100, k = n_i = 10$

	Bootstrap MM	Bootstrap MLE	Bias Correction MM	Bias Correction MLE
Coverage	0.98	0.969	0.933	1
Width	0.234	0.209	0.315	0.998

The parametric bootstrap provides intervals which are too long for either estimation technique, and the bias correction method once again performed poorly.

Simulations with $\mu = 0.1$ and known $M = 4$, but with $k = 30$ observations of $n_i = 30$ trials show that all methods quickly tend to converge in result as the sample size increases.

Table D.5 Empirical and Hierarchical Bayesian Interval Widths and Coverage - $\mu = 0.10, M = 4, n_i = k = 30$

	EB MM	EB MLE	Intermediate Prior Matching $\hat{\mu}_{MM}, \hat{\phi}_{MM}$	Intermediate Prior Matching $\hat{\mu}_{MLE}, \hat{\phi}_{MLE}$	Approx. Morris Matching Prior
Coverage	0.935	0.938	0.948	0.948	0.95
Width	0.15	0.15	0.151	0.151	0.152

Empirical Bayesian coverage is still below the nominal level, but only by a small amount.

Each of the full hierarchical bayesian models produces roughly 95% coverage, with approximately equal interval widths. Maximum likelihood empirical Bayesian intervals are still larger than the Bayesian intervals roughly 11.93% of the time, and method of moments empirical Bayesian intervals are still larger roughly 33.64% of the time.

Table D.6 Non-Bayesian Correction Methods Coverage and Interval Width - $\mu = 0.1, M = 4, k = n_i = 30$

	Bootstrap MM	Bootstrap MLE	Bias Correction MM	Bias Correction MLE
Coverage	0.946	0.948	0.952	0.952
Width	0.153	0.152	0.162	0.161

Bias correction methods tend to work well, as again posterior distributions will tend much more strongly towards a unimodal bell shape, and so correction methods derived for the normal-normal model will perform well.

APPENDIX E. DERIVATION OF WINNING PERCENTAGE ESTIMATOR

Following the simple simulation of Section 4.3, the game is immediately won if the runs scored after nine innings is larger than the runs allowed after nine innings. Alternatively, if the game is tied, the game proceeds one inning at a time. In each extra inning, a win may be produced by either outscoring the opposing team, or by tying and outscoring the additional team in a future inning. This may be written as

$$P(Win) = P\left(\sum_{i=1}^{9} RS_i > \sum_{i=1}^{9} RA_i\right) + P\left(\sum_{i=1}^{9} RS_i = \sum_{i=1}^{9} RA_i\right) \times \\ [P(RS_i > RA_i) + P(RS_i = RA_i)[P(RS_i > RA_i) + P(RS_i = RA_i)[\ldots]]] \quad \text{(E.1)}$$

There is an infinite recursive process on the right hand side, as the probability of a tie is multiplied by the probability of a win in the next inning plus the probability of a tie times the corresponding probability in the next inning.The quantity $P(RS_i = RA_i)$ may be distributed as

$$P(Win) = P\left(\sum_{i=1}^{9} RS_i > \sum_{i=1}^{9} RA_i\right) + P\left(\sum_{i=1}^{9} RS_i = \sum_{i=1}^{9} RA_i\right) \times \\ [P(RS_i > RA_i) + P(RS_i > RA_i)P(RS_i = RA_i) + P(RS_i = RA_i)^2[\ldots]] \quad \text{(E.2)}$$

Inside the $[\ldots]$, the quantity $P(RS_i = RA_i)^2$ must now be distributed. Continuing this distribution process ad infinitum gives an infinite sum $\sum_{k=0}^{\infty} P(RS_i > RA_i)P(RS_i = RA_i)^k$:

$$P(Win) = P\left(\sum_{i=1}^{9} RS_i > \sum_{i=1}^{9} RA_i\right) + P\left(\sum_{i=1}^{9} RS_i = \sum_{i=1}^{9} RA_i\right) \times$$
$$[P(RS_i > RA_i) + P(RS_i > RA_i)P(RS_i = RA_i) + P(RS_i > RA_i)P(RS_i = RA_i)^2 + \ldots] \quad \text{(E.3)}$$

The quantity $P(RS_i > RA_i)$ may then be factored out to give:

$$P(Win) = P\left(\sum_{i=1}^{9} RS_i > \sum_{i=1}^{9} RA_i\right) + P\left(\sum_{i=1}^{9} RS_i = \sum_{i=1}^{9} RA_i\right) P(RS_i > RA_i)\times$$
$$[1 + P(RS_i = RA_i) + P(RS_i = RA_i)^2 + P(RS_i = RA_i)^3 + P(RS_i = RA_i)^4 + \ldots] \quad \text{(E.4)}$$

Since $0 \leq P(RS_i = RA_i) \leq 1$, the quantity $\sum_{k=0}^{\infty} P(RS_i = RA_i)^k$ may be calculated as the sum of an infinite geometric series.

$$P(Win) = P\left(\sum_{i=1}^{9} RS_i > \sum_{i=1}^{9} RA_i\right) + P\left(\sum_{i=1}^{9} RS_i = \sum_{i=1}^{9} RA_i\right) P(RS_i > RA_i)\times$$
$$\left(\frac{1}{1 - P(RS_i = RA_i)}\right) \quad \text{(E.5)}$$

Which then simplifies to equation (4.8)

$$P(Win) = P\left(\sum_{i=1}^{9} RS_i > \sum_{i=1}^{9} RA_i\right) + P\left(\sum_{i=1}^{9} RS_i = \sum_{i=1}^{9} RA_i\right) \left[\frac{P(RS_i > RA_i)}{1 - P(RS_i = RA_i)}\right] \quad \text{(E.6)}$$