

Bayesian Methods for Statistical Inference on the Common Mean from Multiple Data Sources

Z.-Q. John Lu Will Guthrie
National Institute of Standards and Technology
Gaithersburg, MD

Abstract

Combining studies or data from different sources for estimation of some related or common quantities has arisen in a number of applied problems. While there are myriad results on the development of various estimators of the common mean using classical/frequentist statistical approaches, there is lack of unifying rules and understanding of the statistical inference procedures such as confidence intervals or hypothesis testing. Indeed, if the variances are different and unknown, the inferential problem is notoriously difficult, and is related to the well-known Behrens-Fisher problem. This paper will show how Bayesian statistics provides a unifying tool for uncertainty analysis in the most general multi-laboratory and multiple methods problems. Software implementations will be discussed for facilitating popularization of Bayesian methods among practical users.

1. Introduction

The problem of combining results from independent studies for the purpose of estimating or testing a common objective has been important in a number of areas, including biomedical studies and metrology. Naturally, there is a large literature in statistics and metrology that deal with this applied problem. Most of the papers in the statistical literature deal with the problem of deriving point estimators of the common mean.

Consider combining two unbiased estimators x_1 and x_2 that have mean μ and variances σ_1^2, σ_2^2 . If the ratio of the variances is known, then one can use the weighted-mean estimator

$$(1) \quad \tilde{x} = (\mathbf{s}_2^2 x_1 + \mathbf{s}_1^2 x_2) / (\mathbf{s}_1^2 + \mathbf{s}_2^2) = (x_1 / \mathbf{s}_1^2 + x_2 / \mathbf{s}_2^2) / (1/\mathbf{s}_1^2 + 1/\mathbf{s}_2^2)$$

with the associated variance

$$\text{var}(\tilde{x}) = \mathbf{s}_1^2 \mathbf{s}_2^2 / (\mathbf{s}_1^2 + \mathbf{s}_2^2) = (1/\mathbf{s}_1^2 + 1/\mathbf{s}_2^2)^{-1}.$$

Note that both of above quantities are known if the ratio of the variances is known since the only unknown variance cancels out. If the two variances are equal, then \tilde{x} reduces to the simple-mean estimator: $\bar{x} = (x_1 + x_2) / 2$. Indeed, this estimator can perform quite well even in cases where the individual variances are not equal, as its variance, $(\mathbf{s}_1^2 + \mathbf{s}_2^2) / 4$, is smaller than the variance of either individual estimator when $1/3 \leq \mathbf{s}_1^2 / \mathbf{s}_2^2 \leq 3$. In other words, it is better to combine as long as the variances do not differ too much.

In practice, of course, the variances are almost always unknown. It is precisely because the two variances need to be replaced by some data-based estimators that the problem becomes complicated. For example, Graybill and Deal (1959) show that when the variances are unknown, there does not exist any simple linear estimator of the form: $cx_1 + (1-c)x_2$ for $0 \leq c \leq 1$ which uniformly improves on the individual estimators over all possible values of $\mathbf{s}_1^2, \mathbf{s}_2^2$. Thus, the weights in a linearly combined estimator have to be adaptive to the data in order to dominate over the individual estimators. It is natural to replace the unknown variances by their unbiased estimators in (1). For example, if $m_1 s_1^2 / \mathbf{s}_1^2$ is distributed as $\mathbf{c}_{m_1}^2$ and $m_2 s_2^2 / \mathbf{s}_2^2$ is distributed as $\mathbf{c}_{m_2}^2$, then by replacing $\mathbf{s}_1^2, \mathbf{s}_2^2$ by s_1^2, s_2^2 in (1), Graybill and Deal (1959) showed that the resulting combined estimator is an unbiased estimator and has smaller variance than either x_1 or x_2 when the degree of freedoms (DOFs) satisfy $m_1, m_2 \geq 9$.

In addition, in many real problems, there may be biases associated with each estimator x_1, x_2 used to estimate the common mean μ . In metrology, it is common practice to introduce a bias term, often called a type B uncertainty, in addition to the data-based variance, or type A uncertainty. For example, Eberhardt et al (1989) assumed known

bounds for the biases, and Levenson et al (2000) proposed a data-based estimation method on the bounds for the biases. The effect of introducing a bias term in the individual means is that the resulting estimator is a “shrinkage” weighted mean of the individual estimators and the related variance is inflated in order to incorporate the bias, see Section 2.

The two methods/labs problem can also be generalized to multiple methods or labs. If the number of labs or methods bigger than 5 or 10, the model presented by Rukhin and Vangel (1998), which is similar to the one-way random effects ANOVA model, but with heteroscedastic variances, is a natural framework to develop various estimators of the common mean. Rukhin and Vangel proposed a maximum likelihood estimation method, however, which is difficult to compute. They compared their estimates to various simpler approximate estimates such as the Mandel-Paul algorithm. Note that because the within-lab samples (or DOF associated with each individual mean estimator x_i) may be small in many inter-lab studies, and there will usually be many labs involved (more than 10), the shrinkage or smoothing effect on the common mean inference is more pronounced. Although the Rukhin-Vangel approach and similar approaches using random effects ANOVA as in W.G. Cochran’s approach work well in many cases, they do not address the problem of combining a small number of methods or labs.

In some senses, obtaining a point estimate of the common mean may be a relatively simple problem, since the weighted, shrinkage estimator is often close to the simple mean estimator in many cases. It is the inferential uncertainty, such as confidence intervals or hypothesis testing that appears to be the most complicated issue using the classical/frequentist approach. The reason is that, the distribution of any combined estimator of \boldsymbol{m} will involve unknown variances as nuisance parameters, so the standard methods have serious limitations in finding exact confidence intervals. For example, Jordan and Krishnamoorthy (1996) proposed some new methods for deriving exact confidence intervals and reviewed some other methods. In general, the classical frequentist statistical approach to the confidence interval is very complicated and often somewhat ad hoc. The related hypothesis testing is no less difficult.

Indeed, when the variances are unknown and unequal, the inferential problem of the common mean is related to the famous Behrens-Fisher problem. In the two-sample case, we have independent vectors $X_1 = (x_{11}, x_{12}, \dots, x_{1m})$ and $X_2 = (x_{21}, x_{22}, \dots, x_{2n})$ such that $x_{1i} \sim N(\boldsymbol{m}_1, \boldsymbol{d}_1^2)$ and $x_{2i} \sim N(\boldsymbol{m}_2, \boldsymbol{d}_2^2)$ and the sufficient statistics for this problem is: the sample means for each vector, denoted by \bar{x}_1, \bar{x}_2 in accordance with previous notations, with variances $\boldsymbol{s}_1^2 = \boldsymbol{d}_1^2 / m$, and $\boldsymbol{s}_2^2 = \boldsymbol{d}_2^2 / n$, and sample variances for each data vector, s_1^2, s_2^2 . The Behrens-Fisher problem is the inference such as hypothesis testing on the mean difference $\Delta = \boldsymbol{m}_1 - \boldsymbol{m}_2$. Lee (1989), Chapter 5 gave a clear introduction to this area.

Unlike the classical frequentist statistical approaches, in which the three related problems of point estimation, confidence interval, and hypothesis testing are treated separately, the Bayesian statistical approach offers a more unified approach to the inference problems of the common mean from multiple data sets. Under the Bayesian framework, all inference issues, whether point summaries, credible or HPD intervals, or hypothesis testing, will be embodied in the posterior distribution. The difficulty of accounting for uncertainty due to estimation of nuisance parameters can be more easily dealt with through marginalization, or integration of the full posterior distribution. The Bayesian approach is well suited to problems where there are unknown and unequal variances, and where there are biases in the means that need to be estimated from data. When extra information, such as historical data or expert opinion on the method biases, it can be easily incorporated in the estimate of the common mean using Bayesian methods.

2. Bayesian Formulation of the general p Multi-lab/method Problem

Suppose that the parameters of interest from each lab/method are $\boldsymbol{m}_1, \dots, \boldsymbol{m}_p$. The data from p labs or methods can be summarized through data reduction using, e.g. standard statistical sufficiency theory: the statistics for the quantities of interest $\boldsymbol{m}_1, \dots, \boldsymbol{m}_p$ are respectively, x_1, \dots, x_p , with associated variances $\boldsymbol{S}_1^2, \dots, \boldsymbol{S}_p^2$. More precisely, we make the assumption that

$$(2) \quad x_1 \sim N(\mathbf{m}_1, \mathbf{s}_1^2), \dots, x_p \sim N(\mathbf{m}_p, \mathbf{s}_p^2), \text{ mutually independent,}$$

And that there exist Chi-squared random variables s_1^2, \dots, s_p^2 such that

$$(3) \quad m_1 s_1^2 / \mathbf{s}_1^2 \sim \mathbf{c}_{m_1}^2, \dots, m_p s_p^2 / \mathbf{s}_p^2 \sim \mathbf{c}_{m_p}^2$$

where m_1, \dots, m_p are the degrees of freedom (DOFs) of the individual variance estimates from each lab. We also assume that s_1^2, \dots, s_p^2 are independent of x_1, \dots, x_p .

The parameters of interest from each lab/method $\mathbf{m}_1, \dots, \mathbf{m}_p$ are related to the *common* parameter of interest μ through the equations

$$(4) \quad \mathbf{m}_1 = \mathbf{m} + b_1, \dots, \mathbf{m}_p = \mathbf{m} + b_p.$$

Here the b_i 's are lab biases, which are assumed to have normal distribution with zero mean and variance s^2 .

The question now is to obtain estimates for unknown parameters of interest, $\mathbf{m}, \mathbf{s}^2, \mathbf{s}_1^2, \dots, \mathbf{s}_p^2$. By the Bayesian theorem, that the posterior is proportional to the product of likelihood function and the prior distribution on all unknown parameters. By the fact that the lab mean parameters may be integrated out from (2) and (4), the joint posterior distribution of the remaining unknown parameters $[\mathbf{m}, \mathbf{s}^2, \mathbf{s}_1^2, \dots, \mathbf{s}_p^2 | x_1, \dots, x_p; s_1^2, \dots, s_p^2]$ is proportional to:

$$(5) \quad \begin{aligned} & \propto \prod_{i=1}^p [x_i | \mathbf{m}, \mathbf{s}^2, \mathbf{s}_i^2] \cdot \prod_{i=1}^p [s_i^2 | \mathbf{s}_i^2] \cdot [\mathbf{m}] [\mathbf{s}_1^2, \dots, \mathbf{s}_p^2; \mathbf{s}^2] \\ & = \prod_{i=1}^p \exp\left[-\frac{(x_i - \mathbf{m})^2}{2(\mathbf{s}^2 + \mathbf{s}_i^2)}\right] \cdot \prod_{i=1}^p (\mathbf{s}^2 + \mathbf{s}_i^2)^{-1/2} \cdot \\ & \quad \prod_{i=1}^p (\mathbf{s}_i^2)^{-\frac{m_i}{2}} \exp\left[-\frac{m_i s_i^2}{2\mathbf{s}_i^2}\right] \cdot [\mathbf{m}] \prod_{i=1}^p [\mathbf{s}_i^2] [\mathbf{s}^2 | \mathbf{s}_i^2] \end{aligned}$$

where the last term $[\mathbf{m}] \prod_{i=1}^p [\mathbf{s}_i^2] [\mathbf{s}^2 | \mathbf{s}_i^2]$ is from the prior distribution assumption.

The joint posterior distribution (5) embodies all information that is needed to make inference on the unknown parameters $\mathbf{m}, \mathbf{s}^2, \mathbf{s}_1^2, \dots, \mathbf{s}_p^2$. For example, the Bayesian estimators of $\mathbf{m}, \mathbf{s}^2, \mathbf{s}_1^2, \dots, \mathbf{s}_p^2$ based on the mode of the posterior distribution are clearly related to the maximum likelihood estimators of Rukhin and Vangel (1998). But the Bayesian approach also easily handles situations when restrictions on the variance components, such as $\mathbf{s}^2 \geq 0, \mathbf{s}_1^2 > 0, \dots, \mathbf{s}_p^2 > 0$, may be in effect. More importantly, Bayesian approach offers a clear way of incorporating the uncertainty in the unknown nuisance parameters $\mathbf{s}^2, \mathbf{s}_1^2, \dots, \mathbf{s}_p^2$ in the assessment of uncertainty in the common mean inference.

Computation based on (5) is significantly simplified under balanced design assumptions, i.e. when $\mathbf{s}_1^2 = \dots = \mathbf{s}_p^2$, a natural consequence from exchangeability considerations among labs/methods. Then (5) specializes to

$$(6) \quad [\mathbf{m}, \mathbf{s}_1^2, \mathbf{s}^2 | \text{data}] \propto (\mathbf{s}_1^2)^{-\frac{1}{2}v_1} (\mathbf{s}^2 + \mathbf{s}_1^2)^{-\frac{1}{2}p} \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^p m_i s_i^2}{\mathbf{s}_1^2} + \frac{\sum_{i=1}^p (x_i - \bar{x})^2}{\mathbf{s}^2 + \mathbf{s}_1^2} + \frac{p(\bar{x} - \mathbf{m})^2}{\mathbf{s}^2 + \mathbf{s}_1^2}\right]\right\}$$

$$\cdot [\mathbf{m}] [\mathbf{s}^2 | \mathbf{s}_1^2] [\mathbf{s}_1^2], \quad -\infty \leq \mathbf{m} \leq \infty, \quad \mathbf{s}_1^2 > 0, \quad \mathbf{s}^2 > 0,$$

where $v_1 = \sum_{i=1}^p m_i$, $\bar{x} = \frac{1}{p} \sum_{i=1}^p x_i$. To obtain the posterior distribution of the variance components $(\mathbf{s}^2, \mathbf{s}_1^2)$, (6) is integrated over μ yielding

$$(7) \quad [\mathbf{s}_1^2, \mathbf{s}^2 \mid \text{data}] \propto (\mathbf{s}_1^2)^{-\frac{1}{2}v_1} (\mathbf{s}^2 + \mathbf{s}_1^2)^{-\frac{1}{2}(p-1)} \exp \left\{ -\frac{1}{2} \left[\frac{\sum_{i=1}^p m_i s_i^2}{\mathbf{s}_1^2} + \frac{(p-1)s_b^2}{\mathbf{s}^2 + \mathbf{s}_1^2} \right] \right\} [\mathbf{s}^2 \mid \mathbf{s}_1^2] [\mathbf{s}_1^2]$$

$$\mathbf{s}_1^2 > 0, \mathbf{s}^2 > 0,$$

Where $s_b^2 = \frac{1}{p-1} \sum_{i=1}^p (x_i - \bar{x})^2$ is an estimate of the between lab/method variability. Estimation of \mathbf{s}^2 and \mathbf{s}_1^2 should be done jointly, because of the constraints $\mathbf{s}_1^2 > 0, \mathbf{s}^2 > 0$ and the imprecise nature in the inter-lab variance statistics s_b^2 . Box and Tiao (1973, pp. 252-264) amply demonstrated this point. Overall, computation of (7) is fairly straightforward to implement, either numerically, or by Monte Carlo using \mathbf{C}^{-2} distribution representation as used by Box and Tiao. But, because of the heavy-tailed nature of the marginal posterior distribution of \mathbf{s}^2 when p is small, say ≤ 10 , then it is preferable to obtain the posterior distribution of the common mean parameter \mathbf{m} by direct integration of (6) over $\mathbf{s}_1^2 > 0, \mathbf{s}^2 > 0$. Alternatively, one can write the posterior of \mathbf{m} as an expectation of familiar densities over (7):

$$(8) \quad [\mathbf{m} \mid \text{data}] = E_{[\mathbf{s}^2, \mathbf{s}_1^2 \mid \text{data}]} \Phi(\bar{x}, \mathbf{d}^2) [\mathbf{m}],$$

where Φ is the normal density function, and $\mathbf{d}^2 = \frac{1}{p} (\mathbf{s}^2 + \mathbf{s}_1^2)$.

3. Prior Choices

For simplicity, we assume independent priors on all unknown parameters i.e. $[\mathbf{m}, \mathbf{s}^2, \mathbf{s}_1^2, \dots, \mathbf{s}_p^2] = [\mathbf{m}][\mathbf{s}^2] \prod_{i=1}^p [\mathbf{s}_i^2]$. When the number of observations for each method/lab is significant, estimation of individual variance components is relatively easy and is not strongly influenced by prior choices. Furthermore, estimation of \mathbf{m} itself is relatively stable, so we will focus on the prior choices for \mathbf{s}^2 mainly.

Noninformative choices are popular since they require no prior knowledge. Among the potential choices, it turns out that one has to be extremely careful about a common prior choice for variance $\mathbf{p}_1(\mathbf{s}^2) = 1/\mathbf{s}^2$ or the related vague prior using inverse Gamma $\mathbf{p}_1(\mathbf{s}^2) = (\mathbf{s}^2)^{-(e+1)} \exp(-\mathbf{l}/\mathbf{s}^2)$ when e, \mathbf{l} are small. Such vague prior choice for leads to problematic issues for the posterior distribution. Indeed, such a prior choice may lead to an improper or close to improper posterior distribution. If the Laplace noninformative prior $\mathbf{p}_2(\mathbf{s}^2) \equiv 1$ is used, then the posterior distribution is not proper for $p = 2$, and is only a proper density function when $p \geq 3$.

As a general recommendation, we suggest the noninformative prior:

$$(9) \quad \mathbf{p}_3(\mathbf{s}^2) \propto 1/(c + \mathbf{s}^2),$$

where c is some constant to be specified, as also recommended in Box and Tiao (1973, p.372). When all labs have comparable variances: $\mathbf{s}_1^2 = \dots = \mathbf{s}_p^2 = \bar{\mathbf{s}}_a^2$, a natural choice is to take $c = \bar{\mathbf{s}}_a^2$. It can be shown that the posterior of the inter-lab variance is a proper density function, thanks in part to the partially informative prior $\mathbf{p}_3(\mathbf{s}^2) \propto 1/(\bar{\mathbf{s}}_a^2 + \mathbf{s}^2)$. However, the posterior mean (and variance) of \mathbf{s}^2 tends to infinity when $p \leq 2$ (or $p \leq 4$) because the tail of $\mathbf{p}_3(\mathbf{s}^2)$ decays as $(\mathbf{s}^2)^{-(p/2+1)}$ when $\mathbf{s}^2 \rightarrow \infty$. The dependence of this prior choice on the within-lab variance \mathbf{s}_a^2 may be reasonable, considering the fact that estimation of \mathbf{s}^2 , and \mathbf{s}_a^2 is indeed, closely correlated. Box and Tiao (1973, Chapter 3) gave a very detailed discussion of

Bayesian inference using $p_3(\mathbf{s}^2) \propto 1/(c + \mathbf{s}^2)$, and showed that how the Bayesian approach easily overcame some of the complications of negative estimates of \mathbf{s}^2 which have plagued the classical “point estimation” approaches for a long time. Though recent developments of Markov chain Monte Carlo (Robert and Casella 1999) have taken away some of the needs of clever integral manipulations in the traditional Bayesian calculations as discussed in Box and Tiao (1973), we think for the heavy-tailed posterior calculations, the direct numerical integration approach is still much needed, for calculating the posterior density function and HPD credible intervals.

Apart from the noninformative choices, one can certainly consider informative prior choices in order to model expert opinions or historical data. For example, Levenson et al (2000) has used an uniform prior on the between-lab bias. The conjugate prior choice using inverse Gamma is flexible approach for modeling the variance of between-lab bias. That is,

$$(10) \quad \mathbf{s}^2 \sim \mathbf{t}^2 \mathbf{c}_k^{-2}, \text{ or } IG(\frac{1}{2}k, \mathbf{t}^2).$$

Here we use IG to denote the inverse Gamma distribution and $IG(\frac{1}{2}k, \mathbf{t}^2)$ of degree of freedom $\frac{1}{2}k$ and scale \mathbf{t}^2 has the density function of the form:

$$p_4(\mathbf{s}^2) = \frac{(\mathbf{t}^2)^{k/2}}{2^{k/2} \Gamma(k/2)} (\mathbf{s}^2)^{-(\frac{1}{2}k+1)} \exp\left(-\frac{\mathbf{t}^2}{2\mathbf{s}^2}\right), \mathbf{s}^2 > 0.$$

The general setup is easily amenable to a Markov Chain Monte Carlo implementation through (5). To obtain the marginal distribution of \mathbf{m}, \mathbf{s}^2 , one can simply draw random samples of $\mathbf{s}_1^2, \dots, \mathbf{s}_p^2$ from inverse chi-squares, then use Gibbs sampling to generate (dependent) samples through \mathbf{m} and \mathbf{s}^2 alternatively.

4. Examples

The proposed Bayesian methods have been applied to some typical common mean inference problems in the literature, including the Eberhardt et al (1989) example, also studied by Jordan and Krishnamoorthy (1996). Another example is a non-Gaussian example, in which one needs to combine binomial trials using a Bayesian Beta-binomial model in order to account for the between-sample variability.

References

- [1] Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Inference*. Addison-Wesley, Reading, Massachusetts.
- [2] Eberhardt, K.R., Reeve, C.P. and Spiegelman, C.H. (1989). A Minimax Approach to Combining Means, With Practical Examples. *Chemometrics and Intelligent Laboratory Systems*, **5**, 129-148.
- [3] Graybill, F.A. and Deal, R.B. (1959). Combining Unbiased Estimators. *Biometrics*, **15**, 543-550.
- [4] Jordan, S.M. and K. Krishnamoorthy (1996). Exact Confidence Intervals for the Common Mean of Several Normal Populations. *Biometrics*, **52**, 77-86.
- [5] Lee, P. M. (1989). *Bayesian Statistics: An Introduction*. Oxford University Press, New York.
- [6] Levenson, M.S., D.L. Banks, K.R. Eberhardt, L.M. Gill, W.F. Guthrie, H.K. Liu, M.G. Vangel, J.M. Yen, and N.F. Zhang (2000). An Approach to Combining Results from Multiple Methods Motivated by the ISO GUM. *J. Res. Natl. Ints. Stand. Technol.*, **105**, No.4, 571-579.
- [7] Rukhin, A.L. and M.G. Vangel (1998). Estimation of a Common Mean and Weight Means Statistics. *Journal of American Statist. Association*, **93**, No.441, 303-308.