An Application of Gibbs Sampling to Estimation in Meta Analysis: Accounting for Publication Bias

Richard J. Cleary
Mathematics Department
Saint Michael's College
Colchester, Vermont 05439

George Casella
Biometrics Unit
Cornell University
Ithaca, New York 14853

BU-1253-M

July 1994

An Application of Gibbs Sampling to Estimation in Meta Analysis: Accounting for Publication Bias

by

Richard Cleary, Saint Michael's College

and

George Casella, Cornell University

Section 1 - Model and Notation

A researcher reading an article in a scientific journal observes a p-value, x. If the reader is interested in estimating some parameter θ related to x, and knows the distribution $f(x|\theta)$, then any usual statistical procedure (maximum likelihood, Bayesian methods, etc.) could be invoked. If θ is a standardized mean difference, or effect size, in a two sample problem with known and equal variances for the two populations then we can write $\theta=(\mu_2-\mu_1)/\sigma$ where we consider μ_2 the mean of the treatment population, μ_1 the mean of the control and σ is the common standard deviation. We can show (Cleary, 1993) that the density of the p-values is given by $f(x|\theta) = \exp(\Phi^{-1}(1-x)\cdot\sqrt{\frac{n}{2}}\cdot\theta - \frac{n\theta^2}{4})$ where n is the common sample size for the treatment and control groups.

Suppose that the reader suspects that the journal in question publishes all submitted results that are statistically significant at some level α, but is less likely to publish articles with $x > \alpha$. In this case any estimate of θ should take this selection bias into

account. If we assume that non-significant results are published with probability ρ, we can model the selection bias with the weight function $w(x|\rho) = I_{[x<\alpha]} + \rho \cdot I_{[x \geq \alpha]}$. Then the observed value of x is actually coming from the distribution

$$g(x|\theta,\rho) = \frac{f(x|\theta) \cdot w(x|\rho)}{E(w(x|\rho))} \quad . \tag{1.1}$$

Throughout this discussion we take α=.05.

This model and others similar to it have been studied from several points of view. Bayarri and DeGroot (1986,1991) focused attention on publication bias by considering estimates of θ from the above, but only for the cases ρ = 0 and ρ = 1. Iyengar and Greenhouse (1988) discuss joint maximum liklihood estimation of θ and ρ. Cleary (1993) presents a range of values for θ by computing maximum likelihood and Bayes estimates as functions of ρ. The papers of Dear and Begg (1992) and Hedges (1992) propose similar weighted models.

In the following we consider joint Bayes estimation of the parameter of interest, θ, and the selection parameter, ρ. We begin in section 2 with the case of a single observed x, using flat priors on θ and ρ. We apply Gibbs' sampling to determine the posterior distributions of interest, and we examine some of their properties. In section 3 we expand the model to consider an application in meta-analysis with ρ constant for all studies. Then in section 4 we consider a model in which the size of the selection parameter for an individual study is determined by some of the characteristics of the

study itself. These meta-analytic methods also rely on the Gibbs' sampler, as described by Gelfand and Smith (1990), (see also Casella and George (1992) for an introduction) to determine the posteriors. In section 5 we examine the application of our model to a meta-analysis of the effects of coaching on SAT scores reported by DerSimonian and Laird (1983).

Section 2 - Estimation for a Single Study

Our goal in this section is to use Bayes methods to find estimates for θ and ρ in the model (1.1) as above. We note that upon computing the expected value in the denominator, this expression can be written as

$$g(x|\theta,\rho) = \frac{\exp(\Phi^{-1}(1-x)\cdot\sqrt{\frac{n}{2}}\cdot\theta - \frac{n\theta^2}{4})\cdot[I_{[x<\alpha]} + \rho\cdot I_{[x\geq\alpha]}]}{1 - (1-\rho)\cdot\Phi(\Phi^{-1}(1-\alpha) - \sqrt{\frac{n}{2}}\cdot\theta)} \quad . \qquad (2.1)$$

Throughout this section we consider x to be the p-value from a one-sided test of the hypothesis H_0: $\theta = 0$ vs. H_1: $\theta > 0$. Choosing priors $\pi(\theta)$ and $\pi(\rho)$ so that the calculation of posteriors would be analytically tractable seems difficult or impossible. We will use the flat priors $\pi(\theta) \equiv 1$, $-\infty < \theta < \infty$, and $\pi(\rho) \equiv 1$, $0 \leq \rho \leq 1$. These priors are a sensible starting place if we assume there is limited prior knowledge of the sizes of the parameters, and they make a good starting point for exploring the methodology. The conditional posterior distributions can then be written out as

$$\pi(\theta|x,\rho) = \frac{g(x|\theta,\rho)\cdot\pi(\theta)}{\int_{-\infty}^{\infty} g(x|\theta,\rho)\cdot\pi(\theta)\ d\theta}\ , \quad \text{and} \tag{2.2}$$

$$\pi(\rho|x,\theta) = \frac{g(x|\theta,\rho)\cdot\pi(\rho)}{\int_{0}^{1} g(x|\theta,\rho)\cdot\pi(\rho)\ d\rho}\ . \tag{2.3}$$

Before discussing the numerical work necessary to carry out the Gibbs sampling algorithm, we make some observations on the behavior of $\pi(\rho|x,\theta)$ for some special cases.

2.1 - Limiting Behavior of the Posteriors

A reader observing a single significant p-value in a journal could, depending on their attitude about publication bias, interpret that result in two very different ways. If unconcerned with publication bias they would take any given estimates at face value. If certain that publication bias exists they may dismiss the result as a Type I error that found its way into publication. Because the range of reasonable interpretations is so large it is important that any model claiming to shed light on publication bias be well understood in the extreme cases. In this section we consider the limiting behavior of the posterior (2.2).

Looking carefully at the formula for $\pi(\rho|x,\theta)$ in equation 2.3 and substituting $g(x|\theta,\rho)$ from (2.1) while setting $\pi(\rho)=1$ we get

$$\pi(\rho|x,\theta) = \frac{[I_{[x<\alpha]}+\rho\cdot I_{[x\geq\alpha]}]/(1-(1-\rho)\cdot\Phi(\Phi^{-1}(1-\alpha)-\sqrt{\frac{n}{2}}\cdot\theta))}{\int_{0}^{1}[I_{[x<\alpha]}+\rho\cdot I_{[x\geq\alpha]}]/(1-(1-\rho)\cdot\Phi(\Phi^{-1}(1-\alpha)-\sqrt{\frac{n}{2}}\cdot\theta))d\rho} \tag{2.4}$$

Though complicated, this expression can be evaluated exactly by considering separately the cases $x<\alpha$ and $x \geq \alpha$. Using the shorthand

$$A = A(\theta) = \Phi(\Phi^{-1}(1-\alpha) - \sqrt{\tfrac{n}{2}}\cdot\theta),$$

the denominator of $\pi(\rho|x,\theta)$ can be evaluated as follows:

For $x<\alpha$,

$$\int_0^1 \frac{1}{1-(1-\rho)\cdot A}\, d\rho = \frac{-\ln[1 - A]}{A} .$$

For $x \geq \alpha$,

$$\int_0^1 \frac{\rho}{1-(1-\rho)\cdot A}\, d\rho = \frac{A + (1 - A)\cdot\ln(1 - A)}{A^2} .$$

Substituting these expressions back into equation (2.4) gives us an expression for $\pi(\rho|x,\theta)$ which depends on the significance of x. The next two theorems note the asymptotic behavior of $\pi(\rho|x,\theta)$. We first consider the case where $x<\alpha$, which corresponds to the observation of a significant p-value.

Theorem 2.1: For $x<\alpha$, we have

a.) $\lim_{\theta\to-\infty} \pi(\rho|x,\theta) =$

$$\begin{cases} 0 \text{ if } \rho \neq 0 \\ \infty \text{ if } \rho = 0 \end{cases}$$

b.) $\lim_{\theta\to\infty} \pi(\rho|x,\theta) = 1$ for all ρ.

Proof: Evaluations of the limits are straightforward.

These results show that the model behaves sensibly for extreme values of θ. Examining part (a.) in practical terms we reason as follows. We have observed a significant p-value x. If θ is actually very small, say a negative number of large magnitude, this is very unlikely. We would expect almost all results to be non-significant. The fact that the one observed value is significant is strong evidence of publication bias, and sends all of the posterior mass to be concentrated at ρ=0. In (b.) the increasing value of θ makes it very unlikely that a non-significant x would ever be observed. Since any study will be published if it shows a significant result, the selection bias mechanism does not come into play and thus little information about the selection parameter is acquired. The posterior tends to retain the uniform shape of the prior.

We now examine the behavior of $\pi(\rho|x,\theta)$ in the case when $x \geq \alpha$, which corresponds to the observation of a non-significant p-value.

Theorem 2.2: For $x \geq \alpha$, we have

a.) $\lim_{\theta \to -\infty} \pi(\rho|x,\theta) = 1$.

b.) $\lim_{\theta \to \infty} \pi(\rho|x,\theta) = 2\rho$.

Proof: The evaluations of the limits with L'Hopital's Rule are straightforward.

Again these results make sense in light of the model in use. Part (a.) tells us that if the effect size is very small, observation of a single non-significant study does not provide much

information about the selection parameter. The result in (b.) is also not surprising, it indicates that observation of a p-value greater than α when θ is actually large means that it is likely that ρ is quite large. Overall, the effect of these two theorems is to suggest that the model employed is a reasonable one.

2.2 - Implementing the Gibbs Sampler

Now we return to the question of computing numerical estimates for θ and ρ, applying the Gibbs sampling algorithm to the posteriors $\pi(\theta|x,\rho)$ and $\pi(\rho|x,\theta)$. By choosing a starting value for ρ, call it ρ_0, we can iteratively sample between these two conditionals for a given x. More precisely we choose θ_j by sampling at random from the distribution $\pi(\theta|x,\rho_{j-1})$, then choose ρ_j by picking randomly from $\pi(\rho|x,\theta_j)$. This produces two interlocking Markov chains whose stationary distributions are $\pi(\theta|x)$ and $\pi(\rho|x)$. We can then use usual Bayes measures, such as the posterior mean or median, as our estimates of θ and ρ.

Application of the Gibbs sampler depends on being able to sample at random from the distributions $\pi(\theta|x,\rho)$ and $\pi(\rho|x,\theta)$. Due to the complicated expressions for these distributions, this must also be done using another Markov chain method, the Metropolis algorithm (Metropolis et. al. 1953, Tierney 1991).

Table 2.1 presents the posterior means for θ and ρ for various values of x. In each study we assume that control and treatment

sample sizes are both 20. The means are based on 1000 observations each from $\pi(\theta|x)$ and $\pi(\rho|x)$ after ignoring the first 100 values. This waiting period was chosen after testing sets of 100 observations beyond the first 100, and noting no major changes in the means or standard deviations from set to set. Also presented in the table is the MLE for θ calculated at the posterior mean for ρ.

Table 2.1: Posterior means based on 1000 Gibbs sampling iterations.

x	$E(\rho\|x)$	$E(\theta\|x)$	$\hat{\theta}_{E(\rho\|x)}$
.001	.4997	.9301	.9470
.010	.4931	.6470	.6678
.045	.4614	.3997	.4380
.100	.5663	.2889	.3368
.250	.5747	.1802	.1674

The standard deviations of the estimates in table 2.1 range from .0082 to .0098. Individual values of θ and ρ both have standard deviations of approximately .28 throughout the range of the table.

Two features of table 2.1 stand out immediately. First, we see that our mean for ρ varies only slightly from the prior mean of 0.5. This makes sense as the observation of a single study can obviously tell us very little about the selection process. It is worth noting that the appearance of a value that is just barely significant causes the lowest estimate of the selection parameter. This agrees with our intuition.

Secondly we see that the posterior means for θ are in relatively

good agreement with the MLE evaluated at the posterior mean for ρ. This is no surprise as the flat prior Bayes estimate when there is no selection parameter is essentially the likelihood estimate. For the single study case we thus have little reason to prefer the more complex Bayes approach over simply evaluating the MLE for θ when ρ=.5. In the next section we discover that in the meta-analysis case we can begin to learn more about the value of ρ and make joint estimation less problematic.

Section 3 - Gibbs Sampling and Meta-analysis with Constant ρ

We expand the model of the previous section to the meta-analysis case. We begin by examining the case in which the selection parameter ρ is constant across studies. The analysis is very similar to section 2.2. We again choose the flat priors $\pi(\theta) \equiv 1$ and $\pi(\rho) \equiv 1$. Here we assume we have observed the p-values for k independent tests of hypothesis as described previously, we denote these by $\mathbf{x} = (x_1,\ldots,x_k)$. Similarly we let n_i represent the control and treatment sample sizes if they are identical, or we may use m_i to stand for the harmonic mean of the control and treatment sample sizes if these are different.

The joint density of the x_i's can be written as

$$g(\mathbf{x}|\theta,\rho) = \frac{\exp(\sum_1^k(\Phi^{-1}(1-x_i)\sqrt{m_i}\cdot\theta - \frac{m_i\theta^2}{2}) \cdot \prod_1^k(I_{[x_i \le \alpha]} + \rho I_{[x_i>\alpha]})}{\prod_1^k(1 - (1-\rho)\cdot\Phi(\Phi^{-1}(1-\alpha) - \sqrt{m_i}\cdot\theta))} .$$

It is this distribution that we use to find the form of $\pi(\rho|\mathbf{x},\theta)$ and $\pi(\theta|\mathbf{x},\rho)$. In this case the choice of the flat priors means that the conditional posteriors will simply look like $g(\mathbf{x}|\theta,\rho)$ divided by the appropriate marginal. We can then apply the Gibbs sampler as in the previous section.

Table 3.1 contains the results of this procedure carried out for a meta-analysis of k=4 studies, using several $\mathbf{x}$ values. We again print the posterior means for ρ and θ based on 1000 observations from the posteriors $\pi(\rho|\mathbf{x})$ and $\pi(\theta|\mathbf{x})$, and for comparison we provide the MLE for θ evaluated at the mean of the ρ values. In the table we assume each study had sample sizes of n=20 for both control and treatment groups.

Table 3.1: Posterior means based on 1000 Gibbs sampling iterations.

Meta-analysis case.

Case	$\mathbf{x}$	$E(\rho\|x)$	$E(\theta\|x)$	$\hat{\theta}_{E(\rho\|x)}$
1	(.001,.001,.001,.001)	.4672	.8957	.9450
2	(.010,.010,.010,.010)	.3804	.5842	.6398
3	(.045,.040,.035,.030)	.1191	.1408	.3021
4	(.080,.060,.040,.020)	.5513	.4364	.4598
5	(.400,.300,.200,.100)	.6366	.1753	.1902

Notice that the observation of several studies all of which are just barely significant (Case 3) causes the estimate of ρ to be very low. The relatively small size of the posterior means for θ compared to the MLE's evaluated at the corresponding mean ρ is probably due to

the many extremely small values of θ which are found when the simulated value of ρ is very close to zero. Notice how the presence of a pair of studies not significant at the classical cutoff of α=.05 (case 4) causes a substantially larger estimate of θ than a similar set of p-values (case 3) that are all "just barely" significant. Densities for the distributions $\pi(\rho|x)$ and $\pi(\theta|x)$ corresponding to case two of Table 3.1 above are shown in figures 3.1 and 3.2.

Section 4 - A Model for ρ Based on Study Characteristics

We turn now to the problem of performing a Bayesian analysis when the size of ρ varies from study to study. One simple approach in theory, but difficult to implement in practice or computationally, would be to assign a different prior distribution for each ρ_i. A more sensible approach is to focus on what characteristics of the study make us believe it was more or less likely to be published, and then build a model which explicitly includes those characteristics.

Let t_i be the value of some study characteristic. By choosing an appropriate scale of measurement we could use sample size, amount of funding, reputation of journal, previous work of the authors or any other variable we think is important in determining probability of publication. An interesting effort of this type has been carried out for a set of clinical trials by Berlin, Begg and Louis (1989). We consider trying to model the value of ρ_i by assuming the equation

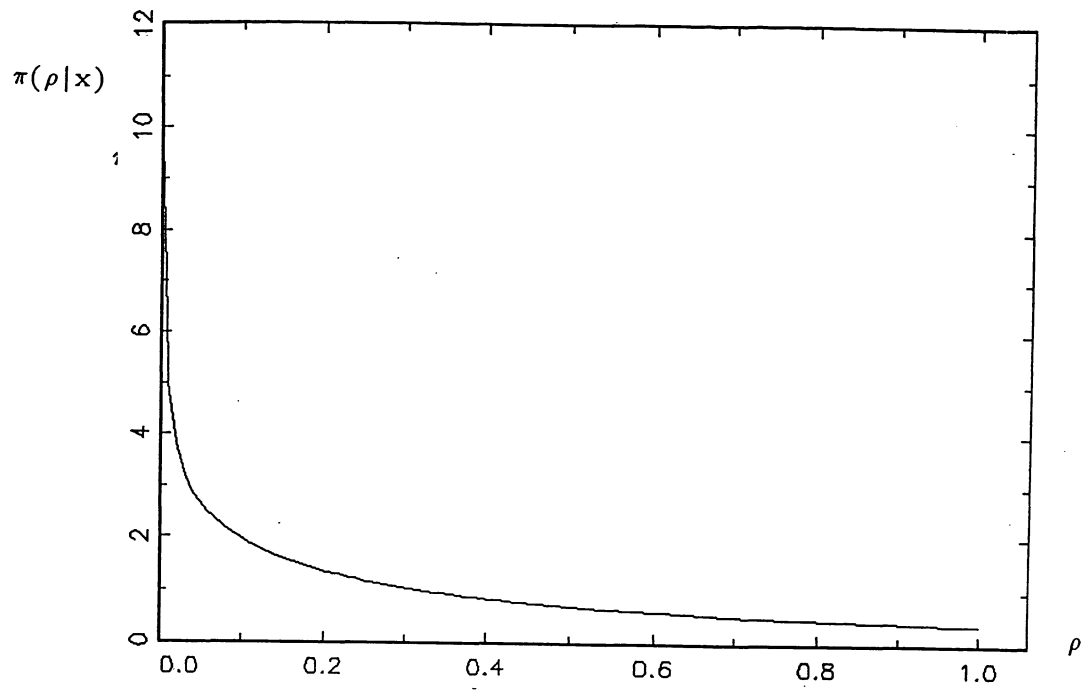

Figure 3.1 - Posterior density of ρ given **x** = (.01,.01,.01,.01)

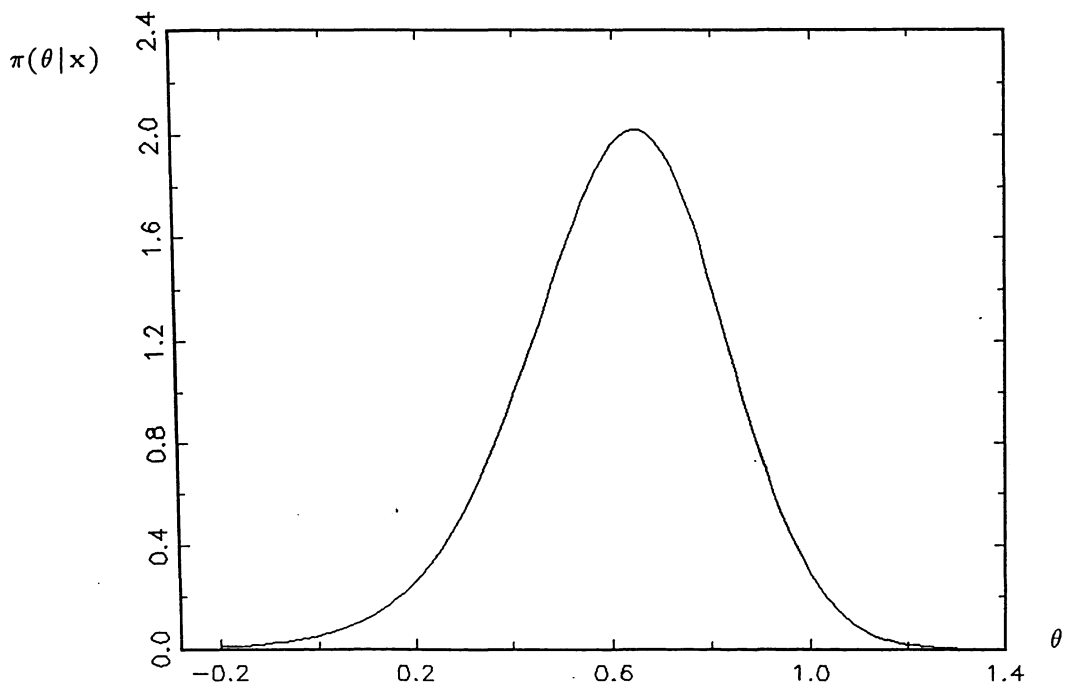

Figure 3.2 - Posterior density of θ given **x** = (.01,.01,.01,.01)

$$\text{logit}(\rho_i) = a + b \cdot t_i \qquad (4.1)$$

holds. It is important to remember that we can not use ordinary logistic regression if all of our observations are from the set of published studies. What we are able to do is put priors on the values of a and b, as well as the effect size θ, and then use Gibbs sampling to find posterior distributions for a, b and θ that are conditional only on x. We then use these conditionals to estimate θ and ρ . This technique can clearly be extended to include more than one study characteristic in the model equation. Alternatively, we could do maximum likelihood estimation of a, b and θ. The reduction in dimension from estimation of k+1 to three parameters might serve to reduce the high variance usually associated with likelihood estimates of selection parameters. We leave this approach for future consideration.

The outline of our estimation plan is as follows:

Step 1.) Determine the study characteristic of most importance in determining publication probability. Choose a measurement scale for this characteristic and calculate t_i for each study.

Step 2.) Choose priors $\pi(a)$, $\pi(b)$ and $\pi(\theta)$ that are appropriate for the problem under consideration.

Step 3.) Write down the forms of the conditional posterior distributions $\pi(a|x,\theta,b)$, $\pi(b|x,\theta,a)$ and $\pi(\theta|x,a,b)$. Choose values a_0 and b_0, and find θ_1 by sampling from $\pi(\theta|x,a_0,b_0)$ using the Metropolis algorithm or other iterative scheme. Continue the Gibbs sampling algorithm, and the resulting values of a, b and θ can eventually be considered as coming from the posteriors $\pi(a|x)$,

$\pi(b|x)$ and $\pi(\theta|x)$.

Step 4.) Using the posterior means from step 3, compute an estimate of ρ_i for each study; and an estimate of the effect size θ. This estimate of θ could be compared to the MLE for θ using the estimated ρ.

We now provide an example in which we choose sample size as the most important predictor of publication probability. Our t_i is taken to be $n_i/50$ where n_i is the common sample size for control and treatment groups. This model considers 50 to be a typical sample size and scales other studies relative to this standard.

The choice of the priors $\pi(a)$ and $\pi(b)$ should be considered carefully. For our example we take $\pi(a)$ to be normal with mean zero and variance nine. The mean of zero corresponds to $\rho=.5$, the variance is chosen to reasonably reflect the logit scale. The prior for b should be a strictly positive distribution since we do not want a large sample size, weighted by a negative value of b, to drag down the probability of publication for a large study. We take $\pi(b)$ to be the exponential distribution with mean and variance 1. We emphasize that this is one of just many possible models. The prior $\pi(\theta)$ is again chosen to be the improper flat prior, $\pi(\theta) \equiv 1$.

A summary of some results from this model for collections of k=2 studies are presented in Table 4.1. The posterior means are computed using 1000 observations after computing and discarding 100 observations. This 100 step "burn in" to the Markov chains seems sufficient as the posterior means examined in groups of 100 show no consistent pattern of increase or decrease. In each case the vector

x represents the p-values from two studies which have sample sizes n_1=20 and n_2=80 respectively. The vector $\boldsymbol{\rho}^*$ in the table represents the values of the ρ_i's from equation (4.1) using the posterior means for a and b. The table includes two estimates of θ, the posterior mean $E(\theta|x)$ from the Gibbs sampling routine, and $\hat{\theta}_{\boldsymbol{\rho}^*}$ which is the MLE for θ evaluated at $\boldsymbol{\rho}^*$.

Table 4.1: Posterior means based on 1000 Gibbs sampling iterations.

Model for ρ_i based on sample size, n_1=20, n_2=80.

Case	x	E(a|x)	E(b|x)	$\boldsymbol{\rho}^*$	E(θ|x)	$\hat{\theta}_{\boldsymbol{\rho}^*}$
1	(.0001,.0001)	-0.440	.574	(.448,.617)	.729	.746
2	(.0100,.0450)	-1.576	.496	(.201,.314)	.254	.271
3	(.0450,.0450)	-1.764	.514	(.174,.281)	.203	.217
4	(.2000,.3000)	-0.609	.531	(.402,.560)	.043	.087
5	(.0450,.3000)	-3.880	.482	(.024,.043)	-.021	-.032

In the first four cases, the components of $\boldsymbol{\rho}^*$ are close in size to the value for a common value of ρ computed by the program of the last section. A dramatic exception occurs in Case 5. Note that this combination gives strong evidence of publication bias based on sample size since a small study with a p-value slightly less than .05 is published, while a larger study with an insignificant value also appears. For the given p-values and sample sizes, the model which assumes a common ρ for each study finds E(ρ|x)=.471; much larger than the estimates of $\boldsymbol{\rho}^*$. This suggests that the model under study is in fact very sensitive to the sample size; future work

could include picking priors for a and b that are more robust.

As usual we see that the application of a more complicated model has resulted in estimates that make sense if the model is indeed correct. If in fact sample size has little to do with the probability of publication, then this model could seriously underestimate that probability. This in turn would lead to an estimate of θ that is too small. We conclude by suggesting that the technique in this section is appropriate if there is strong historical evidence that publication is dependent on other covariates. If there is no such evidence, a statistician performing a meta-analysis might present a range of effect size estimates for different selection models.

Section 5 - Case Study: The Effect of Coaching on SAT Scores

One early and especially thoughtful meta-analysis in the field of education is DerSimonian and Laird (1983). This paper is a review of studies evaluating the value of coaching for students preparing to take the Scholastic Aptitude Test, or SAT. The authors were motivated to undertake the project when two previous reviews of the literature on the subject had reached rather different conclusions as to the effectiveness of coaching and review programs. Both of the previously published reviews had been criticized for their failure to exclude studies of questionable validity, so a further examination of the issue seemed in order.

The problem considered by DerSimonian and Laird exactly fits our description of a two-sample problem with a control group of uncoached students, and a treatment group of students who have received some preparation for the exam. The variable of interest is the mean number of points gained by the treatment group, this can be converted to an effect size by dividing by the standard deviation. Our assumption of known variance, in this case $\sigma^2 = 100^2$, is reasonable since the SAT is designed and thoroughly pre-tested to produce such a result. (The authors and previous reviewers both propose separate analyses for the verbal and mathematics portions of the SAT.) We are also quite clearly interested in the one-sided testing problem, as we unlikely to believe that a coaching program would cause a decrease in scores. Finally, and perhaps most importantly, DerSimonian and Laird mention in the introduction to their methodology that they are choosing to ignore the effects, if any, of selection bias. It will be interesting to see how consideration of this will effect the conclusions of their study.

In this paper we consider only the set of studies that the authors consider most reliable, those in which students were carefully matched or randomized to treatment or control. Table 5.1 provides the data from these studies. Note that the estimated effect size given for each study is the value computed without adjusting for selection bias. The last column represents the estimated value of ρ for each study computed using the model of section 4.

Table 5.1 Sample sizes, estimated effect size without publication bias, p-value, and estimated value of ρ for matched or randomized studies of coaching effectiveness.

i	$n_{t,i}$	$n_{c,i}$	obs. effect (θ_i)	p-value(x_i)	ρ^*
1	45	45	.084	.3446	.558
2	52	52	.110	.2877	.575
3	154	111	.144	.1251	.738
4	239	320	.084	.1635	.918

Using the data in this table, DerSimonian and Laird estimate the effect size, using a simple random effects model, as .101. Using our MLE with $\boldsymbol{\rho}$ =(1,1,1,1) we get essentially the same value, $\hat{\theta}$ = .102. (The slight difference is accounted for because they use the actual variances instead of assuming σ=100. This also accounts for a difference in the size of the reported p-values.) Thus we see that for the case of no selection bias, our model confirms their result. We now analyze the data in table 5.1 using the methods of section 3 and section 4.

Since all four reported studies show a p-value insignificant at α=.05, there is not a great deal of evidence to suggest that there is any sort of strict publication bias at that level. Given the nature of the data and the size of the studies, we consider making the cutoff for certain publication at α=.20. Using this cutoff and applying the flat-prior Bayes approach of section 3 we find $E(\rho|\mathbf{x})$=.538 and $E(\theta|\mathbf{x})$=.052. Graphs of the posterior distributions $\pi(\rho|\mathbf{x})$ and $\pi(\theta|\mathbf{x})$ are attached as figures 5.1 and 5.2.

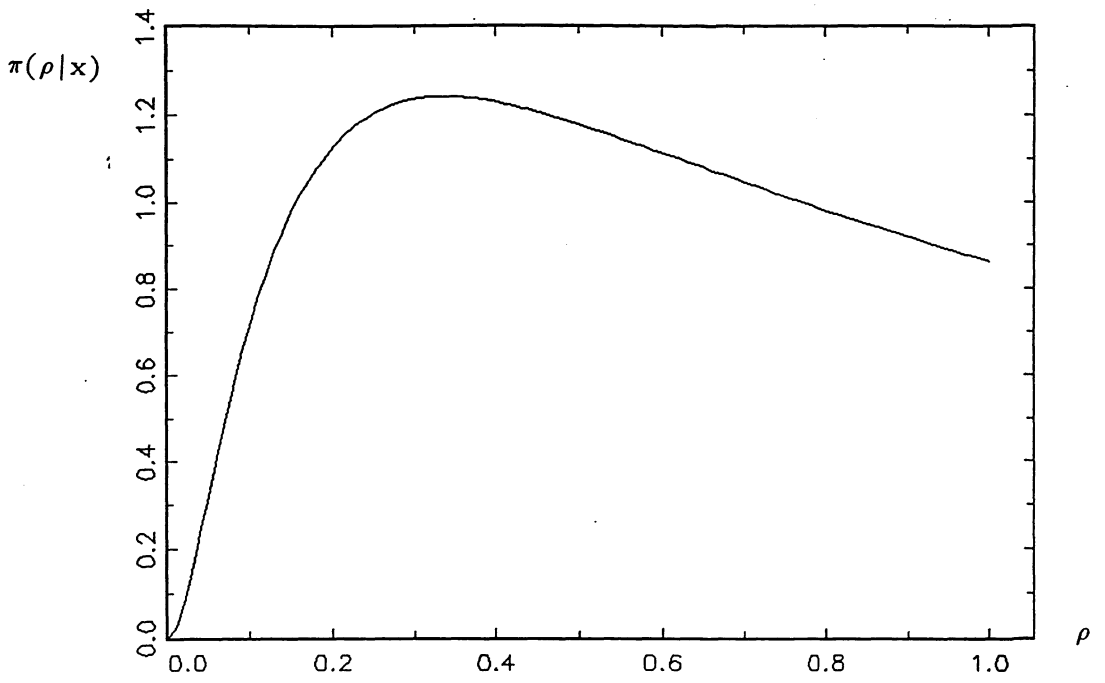

Figure 5.1 - Posterior density of ρ for the SAT data using the flat prior Bayes approach.

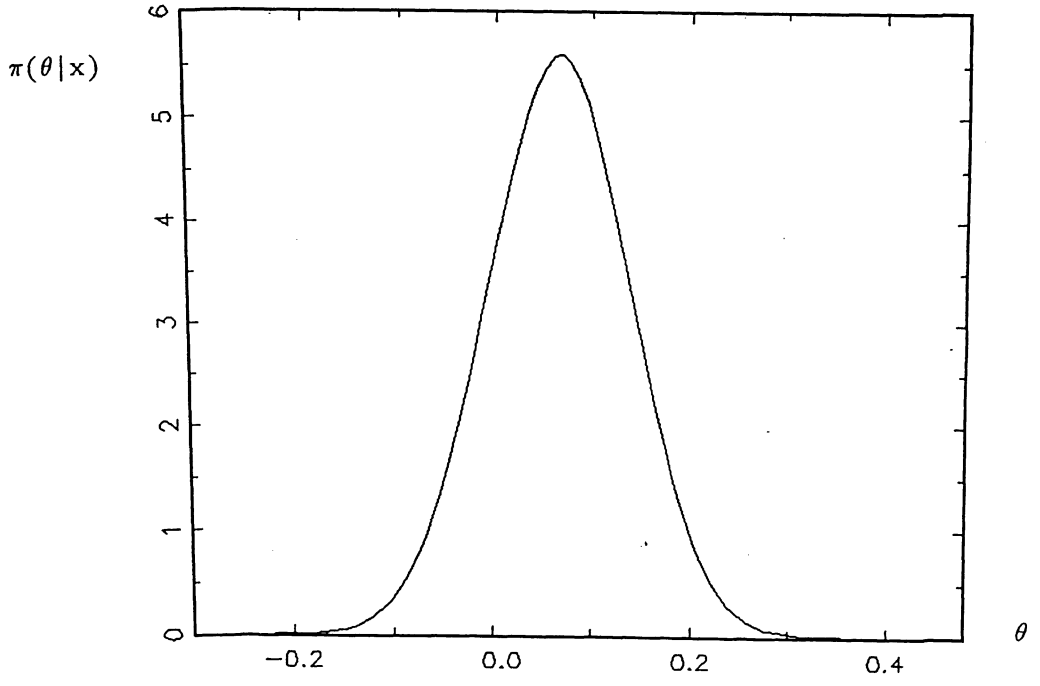

Figure 5.2 - Posterior density of θ for the SAT data using the flat prior Bayes approach.

Several features of the posterior densities are worth noting. In Figure 5.1 we see that the presence of two studies that are not significant at our cutoff of α=.20 shifts the posterior for ρ away from zero. This lack of small values for ρ accounts for the symmetry of the posterior for θ seen in Figure 5.2. (Compare to Figure 3.2 where $\pi(\theta|\mathbf{x})$ is skewed left when the distribution of ρ is concentrated near zero.) While DerSimonian and Laird's overall estimate of $\hat{\theta}$ = .101 is deemed to be just barely significant (at the .05 significance level) in their random effects model, we see in Figure 5.2 that our flat prior Bayes model gives $P(\theta \leq 0|\mathbf{x})$ much larger than .05.

The use of the model presented in section 4 yields a similar result. The estimated values of ρ are provided in Table 5.1 These values were computed using the posterior means $E(a|\mathbf{x})$=-0.195 and $E(b|\mathbf{x})$=.477. The posterior mean of θ is .050 and the distribution is virtually identical to that given in Figure 5.2. Again, consideration of publication bias results in the estimate of the effect size losing its tenuous hold on significance.

DerSimonian and Laird conclude that in the best quality studies there is some very slight evidence that coaching improves verbal SAT scores, but that the size of the effect is so small as to be unimporant practically. Incorporating publication bias into the models makes the case for improvement due to coaching even weaker.

ACKNOWLEDGEMENTS

The authors work on the models above was supported by grants from the National Science Foundation (DMS 91-00839) and the National Security Agency (NO 90F-073). Dr. Cleary thanks the Biometrics Unit at Cornell University for appointments to the 1993 summer session and as a visiting fellow for June 1994.

REFERENCES

Bayarri, M.J. and DeGroot, M. (1986). Information in Selection Models. Technical Report 368, Dept. Statistics, Carnegie-Mellon University.

Bayarri, M.J. and DeGroot, M. (1991). The Analysis of Published Significant Results. Technical Report 91-21, Dept. Statistics, Purdue University.

Berlin, J.A., Begg, C.B. and Louis, T.A. (1989). An Assessment of Publication Bias Using a Sample of Published Clinical Trials. J. Amer. Stat. Assoc. **84**, 381-392.

Casella, G. and George, E.I. (1992). Explaining the Gibbs Sampler. *The American Statistician*. **46**, 167-174.

Cleary, R. (1993) Models for Selection Bias in Meta-Analysis. Ph.D. Dissertation, Biometrics Unit, Cornell University.

Dear, K.B.G. and Begg, C.B. (1992) An Approach for Assessing Publication Bias Prior to Performing a Meta-Analysis. Statistical Science. 7, 237-245.

DerSimonian, R. and Laird, N.M. (1983) Evaluating the Effect of Coaching on SAT Scores. Harvard Educational Review. 53, 1-15.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling Based Approaches to Calculating Marginal Densities. *J. Amer. Stat. Assoc.* 85, 398-409.

Hedges, L.V. (1992) Modeling Publication Selection Effects in Meta-Analysis. Statistical Science. 7, 246-255

Iyengar, S. and Greenhouse, J.B. (1988). Selection Models and the File Drawer Problem. *Statistical Science.* 3, 109-135.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *J. of Chemical Physics.* 21, 1087-1091.

Tierney, L. (1991). Markov Chains for Exploring Posterior Distributions. Technical Report 560, School of Statistics, University of Minnesota.