

# Reasons to prefer “Random Effects” (or Bayesian hierarchical models): a primer on “shrinkage” and the “Stein effect”

Consider analysis of a single variable  $y_{ij}$ , where  $i = 1, \dots, n_j$  indexes observations within units and  $j = 1, \dots, J$  indexes units. Assume the data are balanced (i.e.,  $n_j = n, \forall j = 1, \dots, J$ ). Suppose interests focuses on estimating the unit-specific means,  $\mu_j$ ; e.g., a study of education performance, where  $y$  are test scores and the units are classes, schools, or school districts. To make matters more explicit, lets add the assumptions of (a) conditional independence given  $\mu_j$ , (b) conditional homoskedasticity given  $\mu_j$  and  $\sigma^2$ , and (c) normality, to yield the following model:

$$y_{ij} | \mu_j, \sigma^2 \sim N(\mu_j, \sigma^2)$$

with  $\sigma^2$  treated as known for the time being, so as to focus our attention on the  $\mu_j$ . These unit-specific means are trivial to estimate, in that a conventional estimator like the sample mean,

$$\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{ij}}{n},$$

is easy to compute and can be justified as the least squares estimate of  $\mu_j$  and/or a maximum likelihood estimator under our assumption of normality, and moreover, is an unbiased and consistent estimator of  $\mu_j$ . But as it turns out, there is actually a “better” estimator of  $\mu_j$  available, “better” in the sense of having smaller total *mean square error* across the  $J$  estimates of  $\mu_j$ .

We’ll see that treating the  $\mu_j$  as “random effects”, or (more or less equivalently, in a Bayesian setting), giving the  $\mu_j$  an exchangeable prior density, is a preferable way to proceed. We’ll also see how these ideas generalize to the regression setting, when we’re interested in slightly more complicated models than the one considered here.

## 1 Choosing Among Estimators: the Mean Square Error Criterion

Lets review some standard results from decision theory. Your problem is to select an action  $a \in \mathcal{A}$ . You base your decision on your beliefs about a parameter  $\theta$ , which

indexes possible states of nature, i.e.,  $\theta \in \Theta$ . Often we set  $\mathcal{A} = \Theta$  and the action  $a$  is based on an estimate of  $\theta$ ,  $\hat{\theta}$  (more on that, below).

You incur a loss  $L(\theta, a)$  when you choose action  $a$  when the state of nature is  $\theta$ . For instance, *absolute loss* is defined as

$$L(\theta, a) = |a - \theta|$$

and *squared error loss* is simply

$$L(\theta, a) = (a - \theta)^2.$$

Note that for both of these loss functions, loss is an increasing function of the distance between  $a$  and  $\theta$ .

You possess a technology for estimating  $\theta$ , yielding estimates  $\hat{\theta}$ . An unbiased estimate of  $\theta$  has the property

$$b(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta = 0,$$

where for a classical or frequentist statistician, the expectation  $E(\hat{\theta}) = \int_{\Theta} \hat{\theta} f(\hat{\theta}) d\hat{\theta}$  is computed with respect to the *sampling distribution* of  $\hat{\theta}$ ,  $f(\hat{\theta})$ . Similarly, the variance of  $\hat{\theta}$

$$V(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

is also computed with respect to this sampling distribution. When considering the loss associated with an estimator  $\hat{\theta}$ , we will want to take into consideration this uncertainty arising from variability in repeated sampling. The standard way of doing this is via the *risk function*,

$$R(\theta, \hat{\theta}) = E_{\theta} L(\theta, \hat{\theta})$$

where the notation  $E_{\theta}$  indicates that we are averaging losses over possible values of

$\theta$ . If we have quadratic loss, then the risk function is the *mean square error* of  $\hat{\theta}$ :

$$\begin{aligned}
 \text{MSE}(\theta) &= E_{\theta}(\hat{\theta} - \theta)^2 \\
 &= E_{\theta} [\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\
 &= E_{\theta} [(\hat{\theta} - E(\hat{\theta}))^2 + [E(\hat{\theta}) - \theta]^2] + 2E_{\theta} [(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] \\
 &= E_{\theta} [(\hat{\theta} - E(\hat{\theta}))^2 + [E(\hat{\theta}) - \theta]^2] \\
 &= V(\hat{\theta}) + b(\hat{\theta}, \theta)^2
 \end{aligned}$$

i.e., “MSE equals variance plus squared bias”, noting that

$$E_{\theta} [(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)] = [E(\hat{\theta}) - \theta] E_{\theta}[\hat{\theta} - E(\hat{\theta})] = 0$$

since  $E(\hat{\theta}) - \theta$  is constant with respect to  $\theta$  and so can come in front of the expectation operator, and  $\hat{\theta} - E(\hat{\theta}) = 0$ .

MSE thus gives us a convenient form for evaluating the performance of an estimator, taking into consideration both its bias and its variance. It should be stressed that this link between MSE and the risk of an estimator relies on the use of a quadratic loss function: other loss functions do not admit such a convenient risk function. Nonetheless, it would seem that any sensible consideration of the performance of an estimator ought to consider more than just bias.

## 2 Pooling vs Separating: a Bias-Variance Tradeoff

With these results in hand, we now return to the problem we started with. We have  $J$  means to estimate,  $\mu_1, \dots, \mu_J$ . An unbiased estimate of each one is  $\hat{\mu}_j = n^{-1} \sum_{i=1}^n y_{ij}$ . In some applications the units are small ( $n$  is small) and these estimated means have large variances. In this situation or via an appeal to parsimony we might prefer the pooled estimate, the grand mean

$$\bar{y} = \frac{\sum_{i=1}^n \sum_{j=1}^J y_{ij}}{nJ}$$

which although is a biased estimate of any particular  $\mu_j$ , could well be preferred over the unbiased estimate on the basis of its small sampling variability, on a MSE

criterion. The choice of the pooled estimate over the fixed effects estimates could well be determined by non-statistical criteria, say, if the researcher has a strong commitment to parsimony. But to the extent statistical criteria are relevant to this decision, they are

1. is  $n_j$  large? Note that as  $n_j \rightarrow 0$ , the precision of unit-specific also degrades. In particular, recall that under the assumptions that the data are conditionally iid given  $\mu_j$ , the sampling variability of  $\hat{\mu}_j^{(UP)}$  is estimated as

$$\hat{V}(\hat{\mu}_j^{(UP)}) = \frac{\hat{V}(\mathbf{y}_j)}{n_j} = \frac{\sum_{i=1}^n (y_{ij} - \hat{\mu}_j^{(UP)})^2}{n_j} \times \frac{1}{n_j} = \frac{\sum_{i=1}^n (y_{ij} - \hat{\mu}_j^{(UP)})^2}{n_j^2}.$$

noting that we are using the maximum likelihood estimate of the variance of  $\mathbf{y}_j$ , and “UP” stands for unpooled.

2. is the cross-unit variation in  $y$  small relative to the within-unit variation?; that is, are the unit-specific means  $\mu_j$  tightly clustered around the grand mean  $\mu$ , such that the biases in using  $\bar{y} \equiv \hat{\mu}$  as a *pooled* estimate of the  $\mu_j$  is small. The variance of the group means  $\mu_j$  around the grand mean  $\mu$  is denoted as  $\tau^2$ , which is also known as the *between-group* variance.

To test the pooled estimate against the fixed effects estimator, we can use the *F*-test framework. We can estimate the pooled estimate by simply regressing the  $y_{ij}$  on a constant, or we can estimate the within-group means by regressing  $y_{ij}$  on a series of mutually exclusive and exhaustive dummy variables for group membership. The two regression models being considered are

$$y_{ij} = \mu + \varepsilon_{ij}, \tag{1}$$

versus

$$y_{ij} = \mu_j + \omega_{ij}. \tag{2}$$

and momentarily assume  $\text{var}(\omega_{ij}) = \sigma^2, \forall i, j$ .

We can easily estimate each regression and perform an *F*-test, since the more restrictive/parsimonious model (1) nests as special case of model (2), with the null hypothesis  $H_0 : \mu_1 = \mu_2 = \dots = \mu$ . Alternatively, we might re-parameterize model (2)

with the unit-specific effects as *contrasts* or *differences* about the grand mean  $\mu$ :

$$y_{ij} = \mu + \delta_j + \omega_{ij} \quad (3)$$

where  $\delta_j = \mu_j - \mu$ . We can operationalize this model via the identifying constraint that  $\sum_{i=1}^n \delta_i = 0$ .<sup>1</sup> This alternative parameterization would lead to the joint null hypothesis  $H_0 : \delta_1 = \delta_2 = \dots = \delta_n = 0$  (i.e., there is no across-group variation in the group-specific means). This alternative parameterization also lets us see an important feature of the model: taking variances of both side of equation (3), we have

$$\begin{aligned} \text{var}(y_{it}) &= \text{var}(\mu + \delta_j + \omega_{ij}) \\ &= \text{var}(\mu) + \text{var}(\delta_j) + \text{var}(\omega_{ij}) + \\ &\quad 2\text{cov}(\mu, \delta_j) + 2\text{cov}(\mu, \omega_{ij}) + 2\text{cov}(\delta_j, \omega_{ij}) \\ &= \text{var}(\delta_j) + \text{var}(\omega_{ij}) \\ &= \text{between-unit variation} + \text{within-unit variation} \\ &= \tau^2 + \sigma^2 \end{aligned}$$

There are two main weaknesses to the  $F$ -test approach. One is the homoskedasticity assumption, that the within-group variation ( $\sigma^2$ ) is constant across units. If this condition doesn't hold, then the  $F$ -test for pooling is invalid. In practice, we actually have to test whether  $\sigma_i^2 = \sigma^2, \forall i$  before testing fixed effects against the more restrictive pooled specification.

But second, and more importantly, we can actually do better than having to choose between a pooled model and fixed effects. That is, we can estimate a model with *random effects*, which is effectively a *compromise* between fixed effects and the fully pooled model. The random effects framework also make it reasonably straightforward to deal with the possibility of cross-unit heteroscedasticity.

### 3 Random Effects

Formally, a model with random effects estimates the unit-specific  $\mu_j$  as a *weighted average* of the pooled estimate and the unit-specific estimate  $\bar{y}_j$ . The weights that

---

<sup>1</sup>Equivalently, we could simply regress the  $y_{it}$  on an intercept and dummy variables for  $n - 1$  of the  $n$  units.

contribute to this average are the *precisions* or *inverse variances* of the pooled estimate and the unit-specific estimate. We will see that the pooled estimate and the fixed effects estimates will emerge as special cases of the random effects estimator.

Assume we have normal data, that is

$$y_{ij} \sim N(\mu_j, \sigma^2)$$

but where the unit means are drawn from a distribution themselves, i.e.,

$$\mu_j \sim N(\mu, \tau^2)$$

This second distribution is what a Bayesian would recognize as a prior distribution for  $\mu_j$ , and note that we are putting the same prior over the  $\mu_j$  (i.e., we have no prior knowledge with which to distinguish  $\mu_j$  from  $\mu_k$ ,  $\forall j \neq k$ , in which case the  $\mu_j$  can be said to be *exchangeable* and we ought to make the same probability assignments to them, at least *a priori*). Once we condition on the distribution for the data (the first distribution), we get a posterior distribution for  $\mu_j$ :

$$\mu_j | \mathbf{y} \sim N(\hat{\mu}_j, \sigma^2 + \tau^2)$$

where  $\hat{\mu}_j$  is the random effects estimator of  $\mu_j$  (the mean of the posterior density for  $\mu_j$ ), and is given as

$$\begin{aligned} \hat{\mu}_j &= \frac{\frac{\bar{y}_j}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}} \\ &= (1 - B_j)\bar{y}_j + B_j\mu \end{aligned}$$

where

$$B_j = \frac{\sigma^2}{\sigma^2 + \tau^2}$$

is known as the *shrinkage factor*, since it shows how much the grand mean  $\mu$  contributes to the unit-specific random effects estimate  $\hat{\mu}_j$ . This formulation makes clear that random effects estimates of the unit means are compromises or weighted averages between the biased-but-efficient grand mean and the unbiased-but-inefficient unit mean. Notice that the weights are the ratio of the within-unit variation to the total variation.

*Generalizing the one-way ANOVA model to deal with heteroskedasticity.* Note also that nothing in this formulation constrains  $\sigma^2$  to be constant over the units. That is, in general, we can re-state the model and the results presented above as

$$\begin{aligned}
 y_{ij} &\sim N(\mu_j, \sigma_j^2) \\
 \mu_j &\sim N(\mu, \tau^2) \\
 \mu_j | \mathbf{y} &\sim N(\hat{\mu}_j, V_j) \\
 V_j &= \sigma_j^2 + \tau^2 \\
 \hat{\mu}_j &= (1 - B_j)\bar{y}_j + B_j\mu \\
 B_j &= \frac{\sigma_j^2}{V_j}
 \end{aligned}$$

Finally, note that we also recover an improved estimate of the grand mean from this procedure:

$$\begin{aligned}
 \mu | \tau^2, \mathbf{y} &\sim N(\hat{\mu}, V_\mu) \\
 \hat{\mu} &= \frac{\sum_{j=1}^J W_j \bar{y}_j}{\sum_{j=1}^J W_j} \\
 W_j &= (\sigma_j^2 + \tau^2)^{-1} \\
 V_\mu^{-1} &= \sum_{j=1}^J W_j
 \end{aligned}$$

## 4 Frequentist justification (MSE)

The justification for random effects relies on the fact that the result of combining these two possible estimators of  $\mu_j$  “does better” than either estimator alone. By “does better”, we refer to a mean squared error (MSE) criterion.

A famous result by a Stanford statistician, Charles Stein (1955), runs as follows: if we have  $J \geq 3$  means to estimate (and maintain the homoskedasticity assumptions etc made previously), the estimators

$$\hat{\mu}_j^{(\text{Stein})} = \left( 1 - \frac{(J-2)\sigma^2}{\sum_{j=1}^J n_j \bar{y}_j^2} \right)^+ \bar{y}_j, \quad j = 1, \dots, J,$$

have smaller total mean square error than the unit-specific means  $\bar{y}_j$ . Lindley (1962) proposed an improvement on this estimator, which shifts the unbiased estimates  $\bar{y}_j$  towards the grand mean  $\bar{y}$ :

$$\hat{\mu}_j^{(\text{Lindley})} = \bar{y} + \left( 1 - \frac{(J-3)\sigma^2}{\sum_{j=1}^J n_j(\bar{y} - \bar{y}_j)^2} \right)^+ (\bar{y}_j - \bar{y}) \quad j = 1, \dots, J,$$

which is extremely close to the “random effects” estimator. Efron and Morris (1975) would later popularize the connection between Stein’s result and Bayesian estimators; see 8, below.

To repeat the key idea here, the “semi-pooling” or “shrinkage” estimators (e.g., Stein, Lindley, random effects, empirical Bayes, hierarchical Bayesian, etc) are biased, but make better use of the total information available here to generate superior performance in terms of total mean squared error.

When the within-unit variation  $\sigma_j^2$  is small relative to the total variation,  $\bar{y}_j$  comes with relatively high precision, so we don’t shrink our estimate of  $\mu_j$  very far back towards the grand mean  $\mu$ . But if  $\sigma_j^2$  is large relative to  $\tau^2$ , then we don’t have much precision for  $\bar{y}_j$ , and the resulting estimate will shrink back towards  $\mu_j$ .

Notice that the pooled estimate of the  $\mu_j$  emerges as a special case, with the shrinkage factors  $B_j = 1, \forall j$  if  $\tau^2 = 0$ ; i.e., if there is no between-unit variation then there is no point distinguishing among units! Put differently, in this case the unit-specific means,  $\mu_j$  are drawn from a distribution of means *that has zero variance*, and so  $\mu_1 = \dots = \mu_j = \mu$ .

On the other hand, the shrinkage factors can be zero when the between-unit variation is the only source of variation. That is, there is no within-unit variation ( $\sigma_j^2 = 0, \forall j$ ). In this case the random effects estimates turn out to be the fixed effects estimates.

## 5 “Empirical Bayes”

Random effects are sometimes called “empirical Bayes” estimates. This is because it appears that the posterior for the  $\mu_j$  has come from a “prior” for  $\mu$  given “empirically” by the data itself. This terminology is rather misleading and is fast going out of fashion. Priors are best thought of as not coming from the data (most Bayesians think of that

as cheating, counting the data twice), and it is more appropriate to refer to the random effects model as a “hierarchical model”.

## 6 “Hierarchical Models”

The random effects model is hierarchical in that the way that there are two nested stochastic levels: first,  $y_{ij} \sim N(\mu_j, \sigma_j^2)$ , and then  $\mu_j \sim N(\mu, \tau^2), \forall j = 1, \dots, J$ .

This idea generalizes in two directions. First, we can expand the hierarchy (e.g., student test scores are draws from a distribution around a class mean, the class means can be modeled as draws from a distribution around a district mean, and the district mean can be modeled as draws from a distribution around a state mean, and so on).

Second, we can include covariates at any or all levels of the hierarchy. Thus far this discussion of random effects has been solely in terms of estimating means, but the discussion actually generalizes pretty easily to estimating those means as functions of exogenous variables. Econometricians refer to this type of set-up as “random coefficients” (statisticians tend not to say this, since they seem to adopt the convention that once you shift to a Bayesian framework, all parameters are random); sociologists and education researchers call this setup a “multi-level” model.

## 7 Example 1: Meta Analysis of Aspirin Treatment After Heart Attacks

A simple example appears in the accompanying figure; this is a meta-analysis of six studies of the effects of aspirin in prolonging the lives of victims of heart attack (acute myocardial infarction). There are no covariates. In fact, the data consist of the estimated treatment effect  $y_j$  (the placebo mortality rate minus the aspirin mortality rate) and the precision (inverse variance) with which the treatment effect is estimated in each study (a function of the size of each study),  $\sigma_j^2$ . That is, the data are already aggregated to units (studies), which is fine, since we have no other information on the experiment subjects or the studies. The data are from [Draper et al. \(1992\)](#) and are reproduced in Tables 1 and 2. Note that the large AMIS study finds a negative effect, and tends to dominate any pooled analysis.

Study	Aspirin		Placebo	
	Patients	Mortality (%)	Patients	Mortality (%)
UK-1	615	7.97	624	10.74
CDPA	758	5.80	771	8.30
GAMS	317	8.52	309	10.36
UK-2	832	12.26	850	14.82
PARIS	810	10.49	406	12.81
AMIS	2267	10.85	2257	9.70
Total	5599	9.88	5217	10.73

Table 1: Data from Six Studies Estimating the Effect of Aspirin on Survivorship Following Acute Myocardial Infarction

Study	$y_i$	se	$Z_i$	$p_i$
UK-1	2.77	1.65	1.68	.047
CDPA	2.50	1.31	1.91	.028
GAMS	1.84	2.34	0.79	.216
UK-2	2.56	1.67	1.54	.062
PARIS	2.31	1.98	1.17	.129
AMIS	-1.15	0.90	-1.27	.898
Total	0.86	0.59	1.47	.072

Table 2: Data from Six Studies Estimating the Effect of Aspirin on Survivorship Following Acute Myocardial Infarction

The random effects model for these data is just

$$\begin{aligned} y_j &\sim N(\theta_j, \sigma_j^2) \\ \theta_j &\sim N(\theta, \tau^2) \end{aligned}$$

which yields

$$\theta_j | y_j, \theta, \tau^2 \sim N(\hat{\theta}_j, V_j)$$

where  $\hat{\theta}_j = (1 - B_j)y_j + B_j\theta$  and  $B_j = \sigma_j^2 / (\sigma_j^2 + \tau^2)$ .

This relatively simple problem can set up as a maximum-likelihood problem for the unknown parameters  $\theta$  (the mean of the distribution for the random  $\theta_j$ ) and  $\tau^2$  (the variance of this distribution). We don't estimate the  $\theta_j$  directly, but rather, the parameters of the distribution from which they come; but given these parameters and the observed data, we can form easily expectations (best linear unbiased predictions, BLUPs) for the  $\theta_j$ . The likelihood function over  $\theta$  and  $\tau^2$  given data  $\mathbf{y} = (y_1, \dots, y_J)'$  and  $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_J^2)'$  is

$$L(\theta, \tau^2 | \mathbf{y}, \boldsymbol{\sigma}^2) \propto \prod_{j=1}^J \frac{1}{\sqrt{\sigma_j^2 + \tau^2}} \exp \left[ -\frac{1}{2} \sum_{j=1}^J \frac{(y_j - \theta)^2}{\sigma_j^2 + \tau^2} \right] \quad (4)$$

which is easily maximized to yield  $\hat{\theta}$  and  $\hat{\tau}^2$ . These estimates in turn imply

$$\hat{\theta}_j = (1 - \hat{B}_j)y_j + \hat{B}_j\hat{\theta}$$

where

$$\hat{B}_j = \frac{\sigma_j^2}{\sigma_j^2 + \hat{\tau}^2}.$$

Finally, inference for the (a) estimate of the population mean; (b) study/unit-specific means can be done by recourse to a large  $n$  (asymptotic) argument. That is, as  $J \rightarrow \infty$ ,

$$\begin{aligned} \hat{\theta} &\sim N \left( \theta, \left[ \sum_{j=1}^J \frac{1}{\sigma_j^2 + \hat{\tau}^2} \right]^{-1} \right) \\ \hat{\theta}_j &\sim N \left( \theta_j, \sigma_j^2 (1 - \hat{B}_j) \right), \end{aligned}$$

and so standard errors for the estimates come from the square roots of the variance terms just given.

Draper notes that a problem with maximum-likelihood estimation is that these expressions for the variances of  $\hat{\theta}$  and the  $\hat{\theta}_j$  do not account fully for the uncertainty in  $\tau^2$ , and therefore underestimate the true sampling variances. That is, we compute the variances/standard-errors pretending that  $\tau^2$  is constant. This point is elaborated in [Browne and Draper \(2000\)](#), who conclude that Bayesian simulation is the preferred way to estimate these models.

Figure 5 presents the study specific effect estimates (and the implied 95% confidence intervals), along with the random effects estimates. Note the way that the random effects estimates are “shrunk” in two senses: (a) towards one another (towards the overall effect); (b) and have smaller 95% confidence intervals.

## 8 Example 2: Baseball Batting Averages

In a “classic” pioneering article, Brad Efron and Carl Morris analyzed the batting averages of 18 major league players over their first 45 at bats in the 1970 season. According to [Efron and Morris \(1975, 312\)](#)

The problem is to predict each player’s batting average over the remainder of the season .... [using only the data from the first 45 at bats]. This sample was chosen because we wanted between 30 and 50 at bats to assure a satisfactory approximation of the binomial by the normal distribution while leaving the bulk of at bats to be estimated.

For the uninitiated, the batting average is simply the number of base hits divided by the number of plate appearances, where a base hit is when the batter safely reaches first base after hitting the ball into fair territory. Efron and Morris used an arc-sine transformation to convert the batting averages (proportions) to a quantity that has a normal distribution and unit variance. In a reanalysis of the data, [Casella \(1985\)](#) used a normal model for the proportions themselves, using the binomial variance of the average of the observed batting averages  $\bar{y}(1 - \bar{y})/n$  as the normal variance  $\sigma^2$ ; in these data  $\bar{y} = .265$  and  $n = 45$ , so  $\sigma^2 = .004332 = .0658^2$ .

Here I briefly reanalyze these data with a fully Bayesian model, giving you a glimpse of what Bayesian hierarchical modeling looks like. A conjugate, hierarchical normal

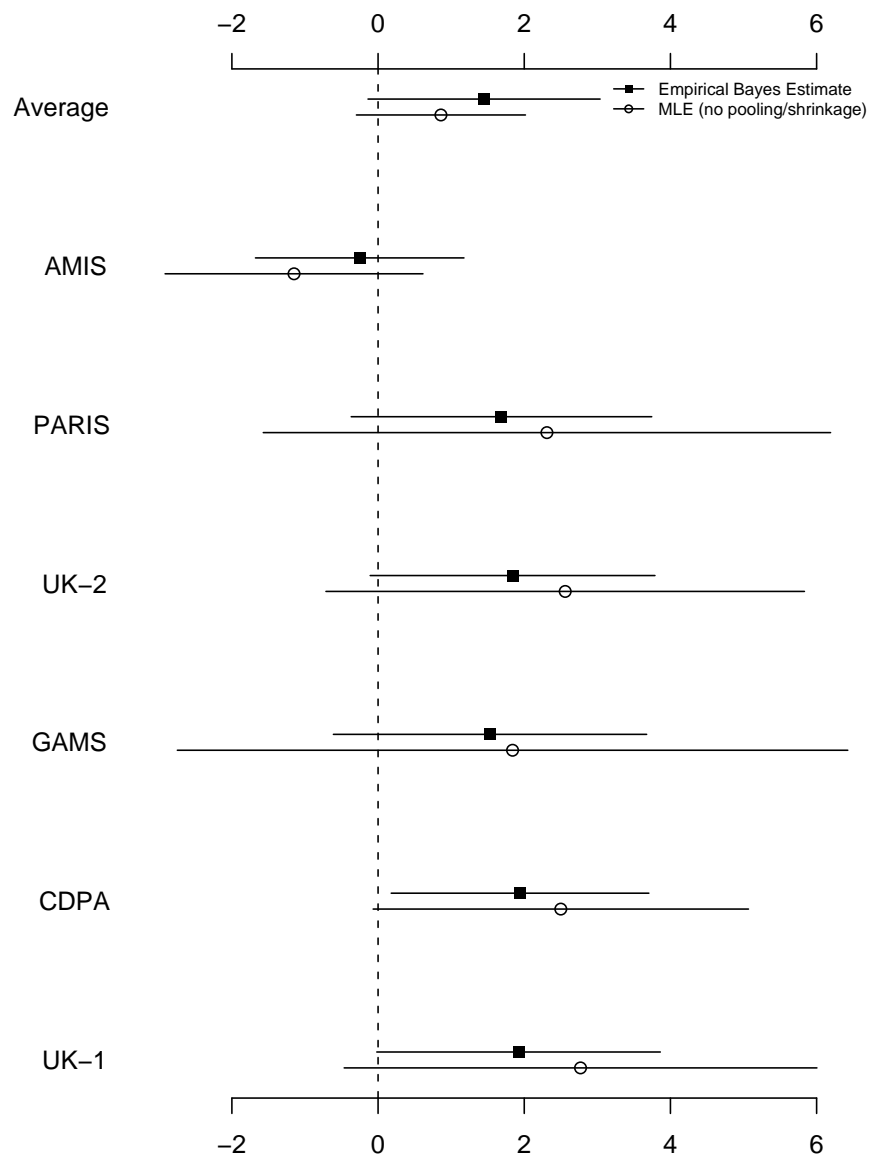


Figure 1: **Meta-Analysis of Effects of Aspirin on Heart-Attack Victims, from Draper et al. (1992, Chapter 1).** In six randomized experiments administered through the 1970s in the USA and Europe, 5,599 heart attack victims were given aspirin, while another 5,217 were given a placebo. The graph shows the treatment effects implied by the MLEs and approximate 95% confidence bounds. The meta-analysis estimates (via random effects) are shown with a solid symbol, have smaller confidence regions than the study-specific estimates, and have “shrunk” towards the (unweighted) mean effect of 0.86. But the random effects/empirical Bayes or “meta-analytic” estimate of the mean treatment effect is 1.45, with a standard error of 0.81. The estimated between-study variance is 1.53.

model for these data is

$$\begin{aligned}
 y_i &\sim N(\theta_i, \sigma^2) \\
 \theta_i &\sim N(\theta_0, \omega^2) \\
 \theta_0 &\sim N(b_0, B_0) \\
 \omega^2 &\sim \text{inverse-Gamma}(v_0/2, v_0\omega_0^2/2)
 \end{aligned}$$

where  $b_0$ ,  $B_0$ ,  $v_0$  and  $\omega_0^2$  are user-specified hyperparameters.

Each observed batting average  $y_i$  is a draw from a normal density with unknown mean  $\theta_i$ . The  $\theta_i$  (the batting average for player  $i$ , which, if the model is correct, we would observe after an arbitrarily long sequence of at bats for each player) are drawn from a normal distribution with an unknown mean and an unknown variance. That is, and consistent with preceding discussion, we regard the players as exchangeable, lacking any prior information to distinguish one from the other. Priors over the unknown mean  $\mu_0$  and unknown variance  $\omega^2$  of the  $\theta_i$  distribution complete the model specification;  $\mu_0$  is the average batting average, while  $\omega^2$  is the variance of the batting averages. We treat these parameters as independent *a priori*, i.e.,  $p(\mu_0, \omega^2) = p(\mu_0)p(\omega^2)$ , with a  $N(b_0, B_0)$  marginal prior density for  $\mu_0$  and an  $\text{inverse-Gamma}(v_0/2, v_0\omega_0^2/2)$  marginal prior density for  $\omega^2$ .

*Priors.* Specifying values for the hyper-parameters completes the prior specification. Baseball batting averages computed over a season (for players with at least 45 at bats) lie in a relatively narrow range: certainly no lower than zero, and almost never above .4 (Ted Williams was the last major league hitter to post a season average above .4, in 1941). The average batting average  $\mu_0$  will lie somewhere in that interval; I guess that the average batting average is unlikely to be greater than .3 and unlikely to be less than .15. The middle of this interval is .225, which I take as the value of  $b_0 = E(\theta_0)$ ; if the interval (.15, .3) corresponds to a 95% prior confidence interval, then  $B_0 = (.15/4)^2 = .00140625$ . For the variance parameter  $\omega^2$ , I use an inverse-Gamma prior density with parameters  $v_0 = 14$  and  $\omega_0^2 = .005$ . I calibrated this prior by noting that the implied marginal prior for  $\theta_i$  is

$$\begin{aligned}
 p(\theta_i) &= \int_{-\infty}^{\infty} \int_0^{\infty} p(\theta_i, \mu_0, \omega^2) d\omega^2 d\mu_0 \\
 &= \int_{-\infty}^{\infty} \int_0^{\infty} p(\theta_i | \mu_0, \omega^2) p(\mu_0) p(\omega^2) d\omega^2 d\mu_0.
 \end{aligned}$$

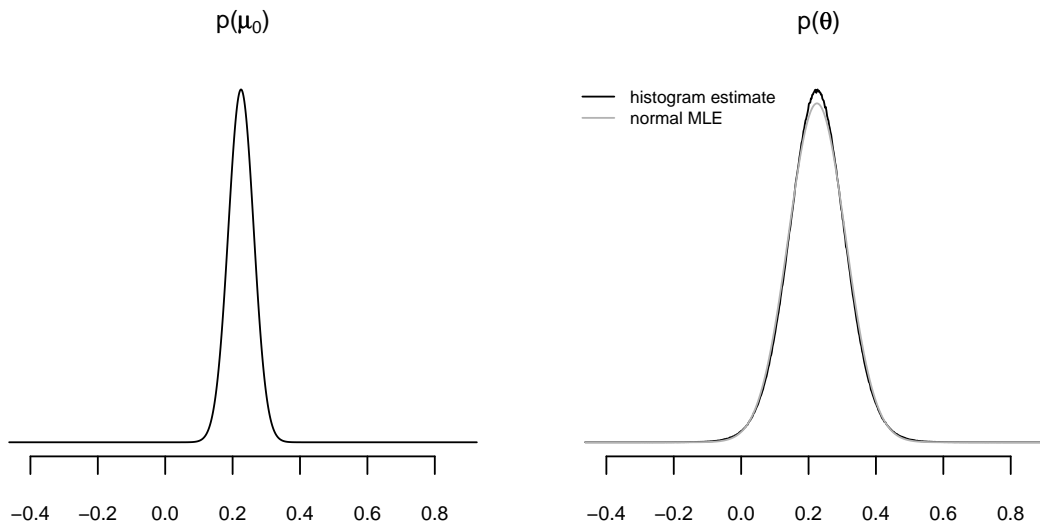


Figure 2: Marginal prior densities of  $\mu_0$  and  $\theta_i$ , baseball averages example.

It is straightforward to sample from this density via the following algorithm: for  $t = 1, \dots, T$ ,

1. sample  $\mu_0^{(t)} \sim N(b_0, B_0)$
2. sample  $\omega^{2(t)} \sim \text{inverse-Gamma}(v_0/2, v_0\omega_0^2/2)$
3. sample  $\theta^{(t)} \sim N(\mu_0^{(t)}, \omega^{2(t)})$

Ten million draws from the marginal prior densities of  $\mu_0$  and  $\theta_i$  appear in Figure 2. The marginal prior for  $\mu_0$  is just a normal density, while the marginal prior for  $\theta$  has slightly heavier tails than the normal, induced by the mixing over the inverse-Gamma density for the between-variance parameter  $\omega^2$ . Very little prior probability is given to impossible, negative batting averages (less than 0.5%); with the normal model for the averages it is impossible to rule out these type of outcomes, although they are extremely unlikely *a priori*. Likewise, ridiculously high batting averages (say, greater than .5) are also effectively zero weight, *a priori*.

*Posterior Density.* The posterior density for this problem is

$$p(\theta_1, \dots, \theta_n, \mu_0, \omega^2 | y_1, \dots, y_n) \propto \underbrace{\prod_{i=1}^n p(y_i | \theta_i)}_{\text{likelihood}} \underbrace{p(\theta_i | \mu_0, \omega^2) p(\mu_0) p(\omega^2)}_{\text{prior}},$$

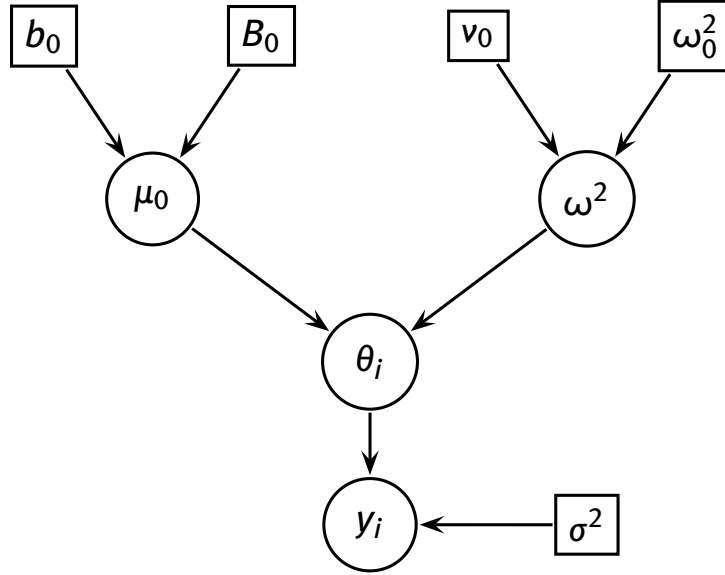


Figure 3: Directed Acyclic Graph, Hierarchical Model for Baseball Batting Averages. Circles denote stochastic quantities (e.g., data or parameters); squares denote known quantities (e.g., user-supplied hyper-parameters).

which looks imposing, but is easy to sample from using the Gibbs sampler. A directed acyclic graph  $\mathcal{G}$  corresponding to this model appears in Figure 3. We can derive the conditional distributions that drive a Gibbs sampler for this problem:

1.  $p(\theta_i | \mathcal{G} \setminus \theta_i)$ . The parents of  $\theta_i$  are the parameters in the hierarchical normal for  $\theta_i$ , the mean  $\mu_0$  and the variance  $\omega^2$ . The only child node of  $\theta_i$  is  $y_i$ , and with the “within” variance  $\sigma^2$  known, the only stochastic parent of  $y_i$  is  $\theta_i$ . Both  $y_i$  and  $\theta_i$  have normal densities, and so  $\theta_i | \mathcal{G} \setminus \theta_i \sim N(\tilde{\theta}_i, V_\theta)$ , where

$$\tilde{\theta}_i = (\mu_0 \omega^{-2} + y_i \sigma^{-2})(\omega^{-2} + \sigma^{-2})^{-1}$$

and  $V_\theta = (\omega^{-2} + \sigma^{-2})^{-1}$ .

2.  $p(\mu_0 | \mathcal{G} \setminus \mu_0)$ . The parents of  $\mu_0$  are the prior hyperparameters  $b_0$  and  $B_0$ . The children of  $\mu_0$  are the  $\theta_i$ ,  $i = 1, \dots, n$ , whose parents are  $\mu_0$  and  $\omega^2$ . Thus

$$p(\mu_0 | \mathcal{G} \setminus \mu_0) \propto p(\mu_0 | b_0, B_0) \times \prod_{i=1}^n p(\theta_i | \mu_0, \omega^2),$$

where all of the densities on the right hand side of this equality are normal densities with known variances, and so

$$\mu_0 | \mathcal{G} \setminus \mu_0 \sim N \left( \frac{b_0 B_0^{-1} + \bar{\theta} \frac{n}{\omega^2}}{B_0^{-1} + \frac{n}{\omega^2}}, \left( B_0^{-1} + \frac{n}{\omega^2} \right)^{-1} \right),$$

where

$$\bar{\theta} = n^{-1} \sum_{i=1}^n \theta_i.$$

3.  $p(\omega^2 | \mathcal{G} \setminus \omega^2)$ . The parents of  $\omega^2$  are the prior hyperparameters  $\nu_0$  and  $\omega_0^2$ . The children of  $\omega^2$  are the  $\theta_i$ ,  $i = 1, \dots, n$ , whose parents are  $\mu_0$  and  $\omega^2$ . Thus

$$p(\omega^2 | \mathcal{G} \setminus \omega^2) \propto p(\omega^2 | \nu_0, \omega_0^2) \times \prod_{i=1}^n p(\theta_i | \mu_0, \omega^2),$$

The prior density  $p(\omega^2 | \nu_0, \omega_0^2)$  is an inverse-Gamma density, while the  $\theta_i$  have normal densities, and so the inverse-Gamma prior over  $\omega^2$  is conjugate with respect to the normal densities over the  $\theta_i$ . Thus

$$\omega^2 | \mathcal{G} \setminus \omega^2 \sim \text{inverse-Gamma} \left( \frac{\nu_0 + n}{2}, \frac{\nu_0 \omega_0^2 + S_\theta}{2} \right)$$

where  $S_\theta = \sum_{i=1}^n (\theta_i - \mu_0)^2$ .

At iteration  $t$  the state of the Gibbs sampler is  $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \dots, \theta_n^{(t)}, \mu_0^{(t)}, \omega^2)^T$ . The sampler makes the transition to  $\boldsymbol{\theta}^{(t+1)}$  by

1. sampling  $\theta_i^{(t+1)}$  from each of the corresponding conditional distributions for the  $\theta_i$ , given  $\mu_0^{(t)}$  and  $\omega^2$ .
2. sampling  $\mu_0^{(t+1)}$  from the conditional distribution for  $\mu_0$ , but given the  $\theta_i^{(t+1)}$  sampled in the previous step.
3. sampling  $\omega^2$  from the conditional distribution for  $\omega^2$ , given the  $\theta_i^{(t+1)}$  and  $\mu_0^{(t+1)}$  sampled in steps 1 and 2.

I initialized the sampler with a random draw from the prior densities defined above, and then let the sampler run for 51,000 iterations, discarding the first 1,000 iterations as burn-in.

*Results.* See Figures 4 and 5. The Bayes estimates (estimated means of the marginal posterior densities of the  $\theta_i$ ) generally lie considerably closer to the actual outcomes than the MLEs. In fact, the Bayesian estimates have a mean square error much smaller than the MLEs; in fact, the MSE of the MLEs is about 2.5 times the MSE of the MLEs. In this case, the shrinkage induced by the hierarchical Bayesian model correctly anticipates the regression to the mean effect: over time, the cross-player variation in the first 45 at bats is attenuated by the subsequent stream of data, that manifests as “reversion to type” (regression to the mean) when we compare the first 45 at bats with the remainder of the season. The hierarchical model does an especially good job of capturing this phenomenon, since it in positing that the players’ averages are drawn from the same distribution (i.e., exchangeability), the model goes some way towards operationalizing the idea that players will be eventually exposed to a roughly same set of competitive scenarios (opposing pitchers, teams, grounds, situations, home and away games, fatigue, etc).

## 9 Additional Reading

Gelman et al. (2004, Chapter 5) provides a good introduction to the more general issue of Bayesian modeling of hierarchical data. The statistician David Draper has done much to popularize the underlying ideas to social-science audiences (e.g., Draper et al., 1992, 1993; Draper, 1995). The idea of shrinkage is quite old, and was first developed outside of the context of random effects or Bayesian modeling (James and Stein, 1960); the result was applied to the random effects case by Lindley and Smith (1972) and a detailed discussion on how the Bayesian analysis is in fact exploiting the James-Stein result appears in Box and Tiao (1973, section 7.2.7). Efron and Morris followed up their now classic 1975 article with an article in *Scientific American* (Efron and Morris, 1977). Reasonably accessible treatments with less of a Bayesian emphasis can also be found in Goldstein (1995) and Kreft and Leeuw (1998).

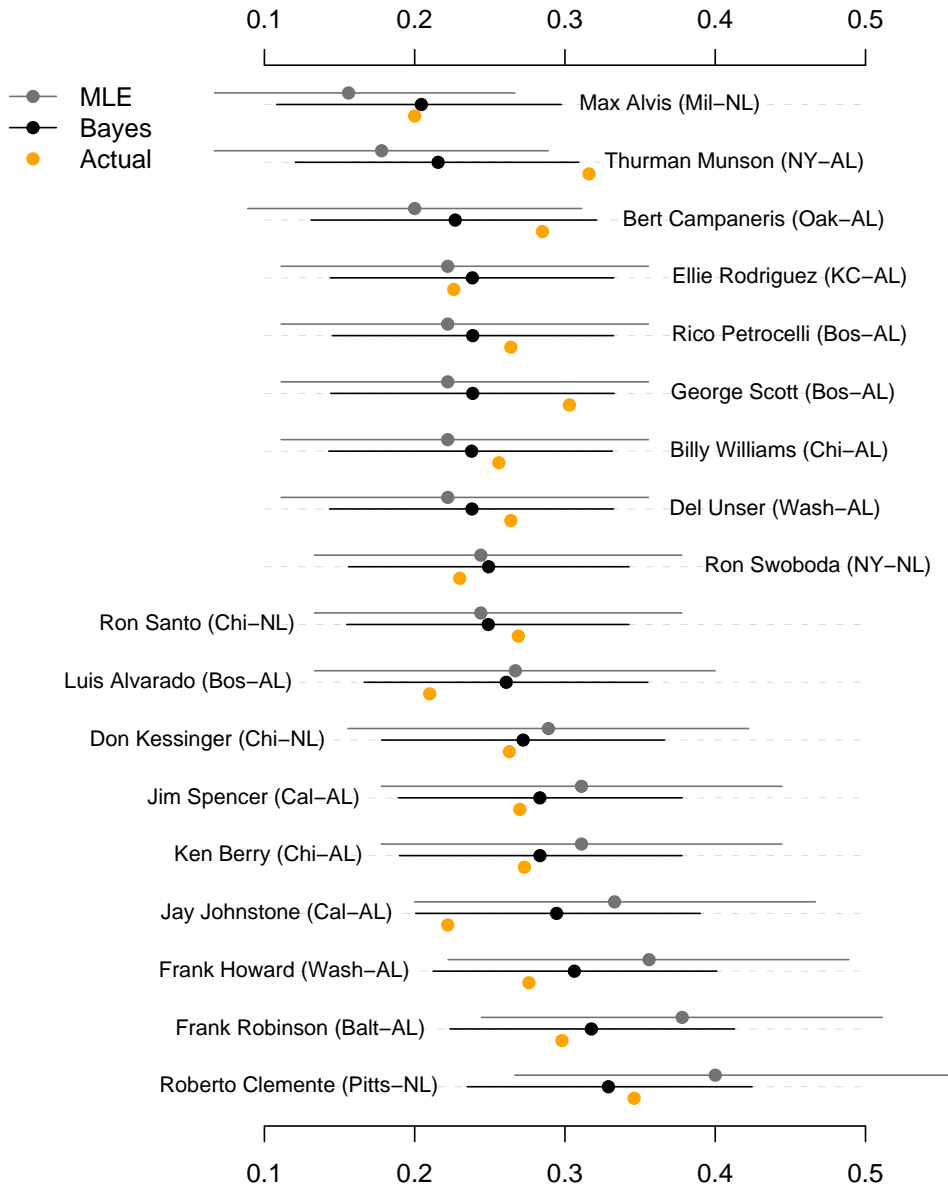


Figure 4: Marginal Posterior Densities, Batting Averages. For each of the 18 players studied, the black horizontal line extends to cover a 95% marginal posterior confidence interval (estimated with the 2.5 and 97.5 percentiles of the 50,000 iterations produced by the Gibbs sampler); the black dot is the mean of the Gibbs sampler output. The gray dot indicates the batting average recorded in the first 45 at bats of the 1970 season (the MLE of the batting average for the rest of the season), and the horizontal lines around it cover a 95% confidence interval. The orange dot is the actual batting average recorded over the remainder of the 1970 season.

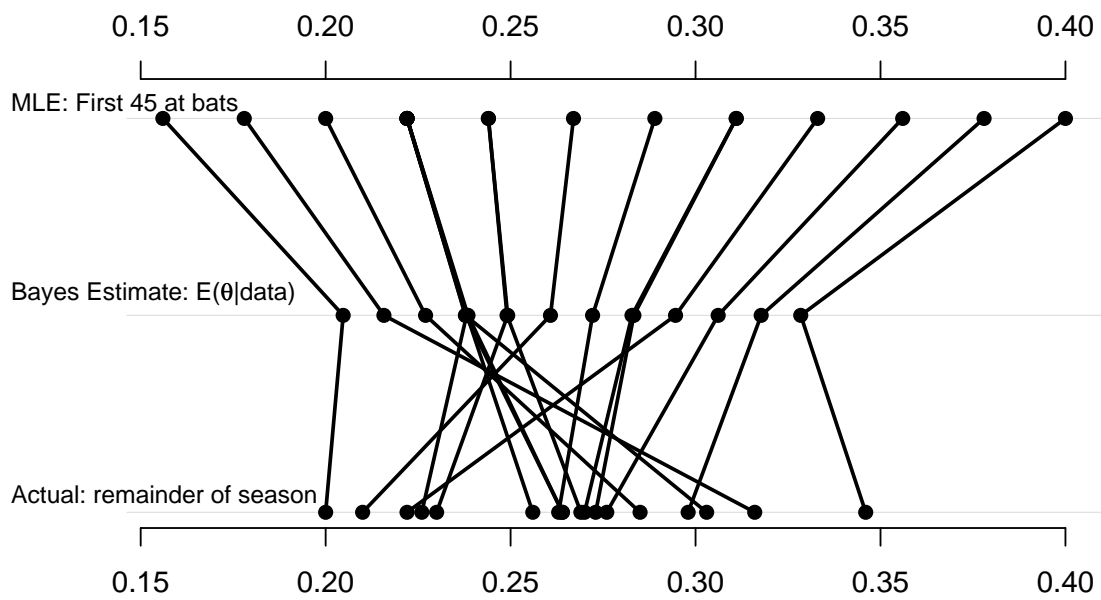


Figure 5: Comparison of MLEs, Bayes Estimates, and Actual Batting Averages. In almost every case, the Bayes estimates (means of the marginal posterior densities) lie closer to the actual batting averages recorded over the remainder of the season than the MLEs. The MLEs have mean squared error 251% that of the Bayes estimates.

## References

- Box, George E. P. and George C. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. New York: Wiley.
- Browne, William J. and David Draper. 2000. "Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models." *Computational Statistics*. To appear.
- Casella, George. 1985. "An Introduction to Empirical Bayes Data Analysis." *The American Statistician* 39:83--87.
- Draper, D. 1995. "Inference and hierarchical modeling in the social sciences." *Journal of Educational and Behavioral Statistics* 20:115--147.
- Draper, D., Donald P. Gaver Jr, Prem K. Goel, Joel B. Greenhouse, Larry V. Hedges, Carl N. Morris, John R. Tucker and Christine M. Waternaux. 1992. *Combining Information: Statistical Issues and Opportunities for Research*. Number 1 in Contemporary Statistics. Alexandria, Virginia: American Statistical Association.
- Draper, D., J. S. Hodges, C. L. Mallows and D. Pregibon. 1993. "Exchangeability and data analysis (with discussion)." *Journal of the Royal Statistical Society, Series A* 156:9--38.
- Efron, Bradley and Carl Morris. 1975. "Data Analysis Using Stein's Estimator and Its Generalizations." *Journal of the American Statistical Association* 70:311--319.
- Efron, Bradley and Carl Morris. 1977. "Stein's paradox in statistics." *Scientific American* 238:119--127.
- Gelman, Andrew, John B. Carlin, Hal S. Stern and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Second ed. Boca Raton, Florida: Chapman and Hall.
- Goldstein, Harvey. 1995. *Multilevel Statistical Models*. Second ed. London: Edward Arnold.
- James, W. and C. Stein. 1960. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman. Vol. I pp. 361--380.

Kreft, Ita and Jan De Leeuw. 1998. *Introducing Multilevel Modeling*. London: Sage.

Lindley, Dennis V. 1962. "Discussion on the article by Stein." *Journal of the Royal Statistical Society, Series B* 24:265--296.

Lindley, Dennis. V. and Adrian F. M. Smith. 1972. "Bayes Estimates for the Linear Model (with discussion)." *Journal of the Royal Statistical Society, Series B* 34:1--41.

Stein, C. 1955. Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*. Vol. 1 Berkeley: University of California Press pp. 197--206.