Empirical Bayes Methods — A Tutorial

by
George Casella
Cornell University and Purdue University

Technical Report #88-18

Department of Statistics
Purdue University

May 1988

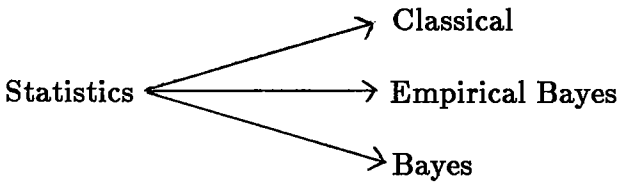
George Casella is Associate Professor, Biometrics Unit, Cornell University, Ithaca, NY. This paper was written while Prof. Casella was on sabbatical leave at Purdue University.

◇

I. What is empirical Bayes (EB)?

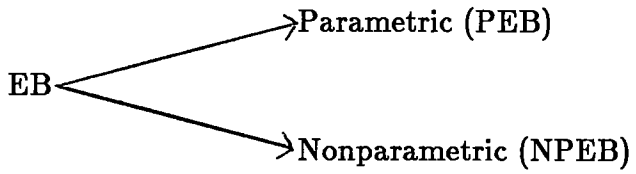
Empirical Bayes is a term that has many meanings, reflecting different approaches to solving problems. It can describe a methodology for both estimation and inference, an important distinction. For the most part, we will be concerned with empirical Bayes estimation.

The general EB approach can be pictured as:



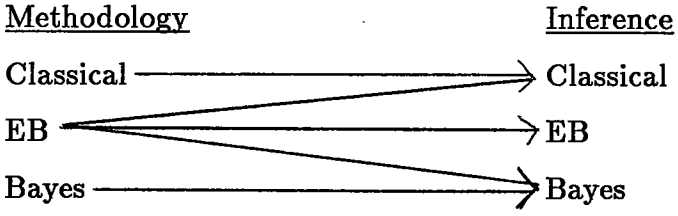
EB methods sit "in between" classical (Neyman-Pearson) and Bayesian statistics, borrowing pieces from each. Although this is necessarily an oversimplification, it serves to put things in perspective.

Within EB methodology, the EB approach can be split into two distinct types



We will concentrate here on PEB techniques. Techniques of NPEB, while quite powerful, are more suitable for large sample analyses, and most properties established for NPEB estimators are large sample properties.

EB can also mean different kinds of inference, but here things get quite involved, as EB inference can borrow pieces from different approaches.



Again, this picture is an oversimplification, but illustrates that the modeling (or estimation) methodology can be quite different from the inferential methodology. EB inference

can be any mix of classical and Bayesian, as can be the inference from any methodology.

II. Statistical Formulations

The general problem is to make an inference about an unknown parameter θ based on observing data $x_1, \ldots, x_n$ according to a sampling distribution $f(x_1, \ldots, x_n|\theta)$.

The assumption of Classical Statistics is that there is a fixed, true value of θ. No prior (subjective) knowledge is available to be used in the estimation or inference process.

In contrast, the Bayes view is that θ is a random variable whose distribution can be quantified with prior (subjective) knowledge. This knowledge can be quantified in a prior distribution, $\pi(\theta)$, and this prior can be updated, using Bayes rule. Using Bayes' rule, the prior $\pi(\theta)$ is updated to the posterior distribution using the formula

$$\pi(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta)\pi(\theta)}{\int f(x_1, \ldots, x_n|\theta)\pi(\theta)d\theta},$$

where $\pi(\theta|x_1, \ldots, x_n)$ is the posterior distribution. The updated distribution, the posterior distribution for θ, is the basis for all inference.

The EB view is a synthesis, in that θ may or may not be random. The Bayes model, in which a hierarchy is used to model. Prior information, that may be useful, is available. This information can be quantified with a family of prior distributions, $\pi(\theta|\tau)$.

Note that the essential difference between PEB and NPEB formulations is the way in which π is treated. In the PEB model, the functional form of π is known, i.e., we have a parametric family of priors. In contrast, the NPEB approach does not assume the functional form of π is known.

As stated before, NPEB is mainly a large sample technique, since many observations are needed to estimate the totally unknown π.

III. A Simple Example

Observe n Bernoulli trials with success probability p. Let $Y = \#$ successes, $Y \sim$ binomial (n, p) with probability mass function

$$f(y|p) = \binom{n}{y} p^y (1-p)^{n-y} = \text{ sampling distribution}$$

First we consider a simple prior distribution on p:

$$\pi(p) = 6p(1-p),$$

which is symmetric about $\frac{1}{2}$. Formally, we can calculate the joint distribution of y and p as

$$\begin{aligned} f(y,p) &= f(y|p)\pi(p) \\ &= 6\binom{n}{y} p^{y+1}(1-p)^{n-y+1} \end{aligned}$$

Also, we can obtain $m(y)$, the marginal distribution of y (unconditional on p), and $\pi(p|y)$, the conditional distribution of p given y, known as the posterior distribution, in the following way.

$$\begin{aligned} m(y) &= \int_0^1 f(y,p)dp \\ &= \int_0^1 6\binom{n}{y} p^{y+1}(1-p)^{n-y+1}dp \\ &= 6\binom{n}{y} \frac{\Gamma(y+2)\Gamma(n-y+2)}{\Gamma(n+4)} \qquad \text{(The beta-binomial distribution)} \end{aligned}$$

and

$$\pi(p|y) = \frac{f(y,p)}{f(y)} = \frac{\Gamma(n+4)}{\Gamma(y+2)\Gamma(n-y+2)} p^{y+1}(1-p)^{n-y+1} \quad \text{(beta distribution)}$$

The posterior distribution plays a most important part in Bayesian statistics, in that it summarizes all information available on the parameter. As noted before, it is an "updated" version of the prior, updated by the data through the use of Bayes rule. In the Bayesian school, all inference comes from the posterior distribution.

A Bayes estimate of p is $E(p|y)$, the mean of the posterior distribution, which is easily calculated as

$$E(p|y) = \frac{y+2}{n+4}.$$

A classical estimate of p, the maximum likelihood estimators $\hat{p}$, is the observed success rate, y/n:

$$\hat{p} = y/n.$$

With some algebra, we have

$$E(p|y) = \frac{y+2}{n+4}$$
$$= \left(\frac{n}{n+4}\right)\hat{p} + \left(1 - \frac{n}{n+4}\right)\left(\frac{1}{2}\right),$$

a weighted average of the classical estimate and the prior mean, with the weights dependent on the sample size. Note that, in general, a Bayes estimate will be a combination of a classical and prior estimates, with weights that reflect the quality of information (that is, variance) of the respective estimators.

If we perform $n = 50$ Bernoulli trials and observe $y = 35$ successes, we get a Classical estimate of p, $\hat{p} = \frac{35}{50} = .7$, and a Bayes estimate of p: $E(p|y) = \frac{50}{54}(.7) + \frac{4}{54}(.5) = .685$.

We can also form Interval Estimates for the classical and Bayes estimators. A simple classical 95% (approx.) confidence interval is given by

$$\hat{p} \pm 2\left(\frac{\hat{p}(1-\hat{p})}{n}\right)^{\frac{1}{2}} = .7 \pm .13 = (.57, .83).$$

A Bayes credible interval can be computed from $\pi(p|y)$. For $y = 35$ we have

$$\pi(p|y = .35) = \frac{\Gamma(54)}{\Gamma(37)\Gamma(17)} p^{36}(1-p)^{16},$$

which is the beta (37,17) distribution. Based on this distribution we can calculate a Bayes 95% credible region: (.56, .80).

Note that the inferences from the classical and Bayesian approach are very different. The 95% Classical Guarantee is that in 95% of all experiments, the procedure $\hat{p} \pm 2(\hat{p}(1-\hat{p})/n)^{\frac{1}{2}}$ will cover the true value of p. In any one realization, however, we do not know if p has been covered. In contrast, the 95% Bayes Guarantee is that the probability is 95% that p lies between .56 and .80. That is, for the *particular data observed*, we specify a 95% coverage probability.

In the PEB approach we start off with a Bayes model, but we specify a family of priors rather than one prior. We have

$$Y|p \sim \text{ binomial } (n, p)$$
$$p|\alpha \sim \frac{\Gamma(2\alpha)}{\Gamma(\alpha)\Gamma(\alpha)}[p(1-p)]^{\alpha-1}$$

where the priors are a family of symmetric Beta distributions

Under this model, we have

$$\pi(p|Y,\alpha) = \text{ Beta } (y+\alpha, n-y+\alpha)$$

and

$$E(p|y,\alpha) = \frac{y+\alpha}{n+2\alpha} = \left(\frac{n}{n+2\alpha}\right)\hat{p} + \left(\frac{2\alpha}{n+2\alpha}\right)\left(\frac{1}{2}\right)$$

FIGURE 3.1 ABOUT HERE

A formal Bayes model requires specifying a value for α, but the PEB model estimates α from the marginal distribution of Y:

$$\begin{aligned} m(y|\alpha) &= \int_0^1 f(Y|p,\alpha)\pi(p|\alpha)dp \\ &= \text{ beta-binomial distribution} \end{aligned}$$

We have a problem with only one observation. We cannot proceed in an EB fashion, since we cannot estimate α. More data is needed in order to be able to estimate α.

Thus, to use the EB approach we need to be able to estimate parameters in the marginal distribution. In some way, we must obtain a sample from the marginal distribution.

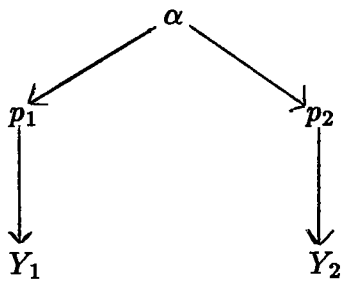
The EB approach is most useful in situations where we have simultaneous estimation problems. Instead of

$$\begin{aligned} Y|p &\sim \text{ binomial } (n,p) \\ p|\alpha &\sim \text{ beta } (\alpha,\alpha) \end{aligned}$$

Consider the situation

$$\begin{aligned} Y_1|p &\sim \text{ binomial } (n,p_1) & \qquad Y_2|p_2 &\sim \text{ binomial } (n,p_2) \\ p_1|\alpha &\sim \text{ beta } (\alpha,\alpha) & \qquad p_2|\alpha &\sim \text{ beta } (\alpha,\alpha) \end{aligned}$$

Pictorially, we can visualize the situation as,



We assume that the two problems are tied together by the underlying common distribution of the p's, but there are some differences in the p_i's, the parameters of main interest.

Suppose now we observe $y_1 = 35$, $y_2 = 27$ according to the model

$$Y_i|p_i \sim \text{ binomial } (n, p_i)$$
$$p_i \sim \text{ beta } (\alpha, \alpha)$$

The Bayes estimate of p_i is, as before,

$$E(p_i|y_i, \alpha) = \left(\frac{n}{n+2\alpha}\right)\left(\frac{y_i+\alpha}{n+2\alpha}\right)\hat{p}_i + \left(\frac{2\alpha}{n+2\alpha}\right)\left(\frac{1}{2}\right)$$

and, marginally, the Y_i's are independent variables with a beta-binomial distribution having

$$EY_i = \frac{n}{2}, \text{ Var } Y_i = \frac{n}{4}\frac{(n+2\alpha)}{2\alpha+1}.$$

Using the method of moments (equating VarY_i with its estimate) we obtain $\hat{\alpha} = 15.205$, and EB estimates

$$\begin{aligned} E(p_i|y_1, y_2, \hat{\alpha}) &= \left(\frac{n}{n+2\hat{\alpha}}\right)\hat{p}_i + \left(\frac{2\hat{\alpha}}{n+2\hat{\alpha}}\right)\left(\frac{1}{2}\right) \\ &= .622\left(\frac{35}{50}\right) + (1-.622)\left(\frac{1}{2}\right) \\ &= .624. \end{aligned}$$

From the Bayes Model we can also calculate

$$\text{Var}(p|y, \alpha) = \frac{(y+\alpha)(n-y+\alpha)}{(n+1\alpha+1)(n+2\alpha)^2} \qquad \begin{pmatrix} \text{from the posterior Beta} \\ (y+\alpha, n-y+\alpha) \text{ distribution} \end{pmatrix}$$

and substituting $n = 50$, $y = 35$, $\hat{\alpha} = 15.205$, the estimated variance is .003 with SD = .054.

To compare, for $y = 35$

	Estimate	SD (posterior)
Classical	.7	.07
Bayes	.685	.063
EB	.624	.054

Note that the EB estimate gives more weight to the symmetric prior than the Bayes estimate.

Alternatively, we can construct an EB interval estimate by looking at the posterior $\pi(p|y, \hat{\alpha})$. We have

$$\begin{aligned} \pi(p|y,\hat{\alpha}) &= \text{Beta } (y+\hat{\alpha}, n-y+\hat{\alpha}) \\ &= \text{Beta } (50.205, 30.205) \quad \text{for } y = 35 \\ &\qquad\qquad\qquad\qquad\qquad\qquad \hat{\alpha} = 15.205, \end{aligned}$$

giving an approximate 95% Credible region = (.52, .73).

These EB standard errors, and hence EB confidence intervals, are very optimistic. Obtaining estimates of error (in this naive way) by substituting data-based values, almost always leads to underestimates of variance. More sophisticated techniques are available, however, that lead to better, more conservative, error estimates. (For example, Laird and Louis (1987), Angers (1987). Some other approaches are given in Section VIII.)

FIGURE 3.2 ABOUT HERE

IV. Some Empirical Bayes Methods in ANOVA

A standard randomized complete block analysis of variance was run to see the effect of linseed oil or digestibility of food. The data are from Hsu (1982).

	Steer Group (Block)						
Treatment	1	2	3	4	5	6	Mean
1	86.5	74.5	68.8	79.9	78.2	86.8	79.1
2	78.2	76.9	67.8	74.2	72.5	76.5	74.4
3	74.7	72.3	72.7	76.3	75.8	76.1	74.7
4	72.9	76.9	64.7	73.2	73.2	73.2	72.4
5	70.8	73.5	67.2	74.5	71.5	70.4	71.3

Here the response is coefficient of digestibility (Y_i), and the treatments are 3 kg/day of hay to which are added increasing amounts of linseed oil meal (approximately 1, 2, 3, 4, 5 kg/animal/day).

In the following analysis we assume that there is no Block $\times$ Treatment interaction. The usual ANOVA model for the treatment means is

$$E\overline{Y}_i = \theta_i$$

One might be tempted now to hypothesize a common prior on θ_i:

$$\theta_i \sim \pi(\theta),$$

(that is, the θ_i's have the same underlying distribution) and construct an EB estimator that uses this information. For example, we can have the model

$$Y_i|\theta_i \sim n(\theta_i, \sigma^2)$$

$$\theta_i|\tau^2 \sim n(0, \tau^2).$$

Or, more generally,

$$Y_i|\theta_i \sim n(\theta_i, \sigma^2)$$

$$\theta_i|r, \tau^2 \sim n(r, \tau^2)$$

$$r \sim \text{ uniform } (-\infty, \infty)$$

Note that the second formulation can be thought of as "shrinking toward the null hypothesis H_0: $\theta_1 = \theta_2 = \cdots = \theta_5$."

Since the experimenter has administered different treatments with the belief that different digestibility will be the result, shrinking toward H_0: $\theta_1 = \cdots = \theta_5$, in this case, does not seem justified. That is, there is no reason to suspect that the diets are equivalent, so shrinking toward H_0 is not justified. It is much better to shrink toward a hypothesis (sub-model) that you believe is true. For example, consider the hypothesis (sub-model)

$$H_0: \quad \theta_i = \alpha + \beta X_i \qquad \alpha, \beta \text{ unspecified},$$

that is, the response exhibits a linear trend, where X_i = amount of linseed oil meal. This trend seems reasonable, given the experiment, since it is reasonable to suspect that increasing X_i will decrease θ_i (linseed oil is difficult to digest!).

We can incorporate this submodel by specifying the Bayes model

$$Y_i|\theta_i \sim n(\theta_i, \sigma^2)$$
$$\theta_i|\alpha, \beta \sim n(\alpha + \beta X_i, \tau^2).$$

Yielding the Bayes estimate of θ_i

$$E(\theta_i|Y_i, \alpha, \beta) = \left(\frac{\tau^2}{\sigma^2 + \tau^2}\right)(\alpha + \beta X_i) + \left(\frac{\sigma^2}{\sigma^2 + \tau^2}\right) Y_i.$$

Marginally, it can be shown that

$$Y_i \sim n(\alpha + \beta X_i, \sigma^2 + \tau^2).$$

As before, we use the marginal distribution of Y to estimate the unknown prior parameters. This is done in two steps.

1. Regressing Y_i on X_i gives estimates for α and β. (standard simple linear regression)
2. Standard normal theory gives us an estimate for $\sigma^2/(\sigma^2+\tau^2)$, details will be presented later.

Coding $X_i = i$, we get $\hat{\theta}_i = 79.66 - 1.76X_i$ with $r^2 = .87$. To summarize, we have the following:

Estimates of θ_i

Treatment	1	2	3	4	5
$\overline{Y}_i$	79.1	74.4	74.7	72.4	71.3
$\hat{\theta}_i$	77.9	76.1	74.4	72.6	70.9
EB	78.9	74.9	74.6	72.5	71.2

The EB estimate is given by

$$EB_i = .29\ \theta_i + .71\ \overline{Y}_i.$$

Note that even though we are not shrinking very much, the shrinkage is strong enough to produce a linear trend in the EB estimates — see the graph for linseed = 2 and 3. The original data are not monotone, but the EB estimates are.

FIGURE 4.1 ABOUT HERE

If we had chosen as our vague prior specification the 'usual' ANOVA null hypothesis H_0: $\theta_1 = \cdots \theta_5$ (which is obviously not a good choice here, and generally is not a good choice), the resulting EB estimate of θ_i would have been

$$.078\ \overline{Y} + .922\ \overline{Y}_i,$$

$\overline{Y}$ ← grand mean

showing that when the prior specification is incorrect, the EB estimate merely collapses back to the cell means.

The usual standard errors and confidence intervals can be used with the EB estimates, and generally the error you will be making is that the intervals will be too wide. For example, the usual 90% t-interval centered at an EB estimate will most often have coverage probability greater than 90%. There is a small chance that the interval will have coverage probability less than 90%, but this is slight enough not to cause worry. So attaching the usual standard errors to the EB estimates is a simple, conservative tactic.

Applying the usual Scheffé procedure, a 90% simultaneous interval for pairwise difference is

$$(\overline{Y}_i - \overline{Y}_j) - 5.251 \leq \theta_i - \theta_j \leq (\overline{Y}_i - \overline{Y}_j) + 5.251$$

that is, with probability 90%, this inequality is specified for all i and j. Using the EB estimates, we can shorten this interval to

$$(EB_i - EB_j) - 5.066 \leq \theta_i - \theta_j \leq (EB_i - EB_j) + 5.066.$$

Obviously, not earthshaking improvement, but reasonable considering p is small (effectively, $p = 3$ because of the restrictions). The improvement becomes more substantial as either p or ν increases, with a 20% decrease in radius a typical reduction for moderately large $p(\geq 10)$. (Assuming, of course, that the prior input is approximately correct.) This topic, of improving on the Scheffé procedure, is treated by Casella and Hwang (1987).

V. Some Other Examples

1. DuMouchel and Harris (1983), investigated interspecies extrapolation of dose-response experiments. They used the model

$$Y_{ij} = \theta_{ij} + \varepsilon_{ij}$$
$$\theta_{ij} = \mu + \alpha_i + \gamma_j + \delta_{ij},$$

where

$y_{ij} =$ observed dose-response slope (log), of species i exposed to environmental agent j

$\theta_{ij} =$ true dose-response slopes

$\mu =$ overall mean

$\alpha_i =$ species-specific effect

$\gamma_j =$ agent-specific effect

Note that the model for the data is a standard "cell means" model, while the submodel is one of "no interaction." It is this feature of the submodel that allows for extrapolation.

A schematic diagram of the (abridged) data is

	Agent					
Species	Roofing Tar Emissions	Coke Oven Emissions	Diesel Engine Emissions	Gas Engine Emissions	Benzo Pyrene	Cigarette Smoke
Human	X	X	O	O	O	X
Mice	X	X	X	X	X	X
Hamster	X	X	X	X	X	X

where X = Data Present and O = Data Absent.

The goals of Du Mochel and Harris were to

1. Provide estimates for cells with no data.
2. Improve Precision of estimates (using posterior standard deviation (SD))

Both goals were to be accomplished by modeling the data as having common underlying structure, and borrowing "ensemble strength" to help improve estimates.

A portion of their results, relating lung cancer risk in humans, is summarized below.

	Estimate (log slope)	Posterior SD	
Roofing Tar			
Orig. Data	.50	1.41	Estimates change a lot because of high SD of original estimate
Bayes	.12	1.02	
EB	.12	1.01	
MLE	−.01	.70	
Coke Oven			
Orig. Data	1.48	.34	Estimates do not change much because of small SD of original estimate
Bayes	1.38	.33	
EB	1.38	.33	
MLE	1.30	.31	
Diesel Engine			
Bayes	−.46	1.45	These values are extrapolated from the analysis. There is no data on humans exposed to diesel engine fumes.
EB	−.46	1.40	
MLE	−.57	.80	

2. Rubin (1980) Law School Validity Study

The object is to predict 1st year final grade average (FGA) of law school students using the equation

$$\text{FGA} = \alpha\ (\text{LSAT}) + \beta\ (\text{UGPA}) + \gamma,$$

where

LSAT = Law School Aptitude Test

UGPA = Undergrad. grade average

The past technique used each of 82 Schools in its own prediction equation, not considering any data from the other law schools. Rubin uses EB to model the law school 'ensemble' and improve estimates

Model:

$$\hat{\beta}_i \sim (\beta_i, \sigma_i^2) \qquad i = 1, \cdots, 82$$
$$\beta_i \sim n(\beta, \tau^2)$$

where

$\hat{\beta}_i =$ individual least squares (LS) estimates of FGA from each law school.

As before, the Bayes estimate for law school i is the posterior mean

$$E(\beta_i|\hat{\beta}_i) = \frac{\tau^2}{\tau^2 + \sigma_i^2}\ \beta + \frac{\sigma_i^2}{\tau^2 + \sigma_i^2}\hat{\beta}_i$$

$$\uparrow$$

$$\text{EB replaces this with } \overline{\beta} = \sum_1^{82} \hat{\beta}_i/82$$

There are complications due to the allowance of unequal variances for each school, details of which we will not go into here. Rubin uses the EM algorithm to obtain EB estimate of $\sigma_i^2/(\sigma_i^2 + \tau^2)$, and then uses these estimates in the above equation. A summary of his results is

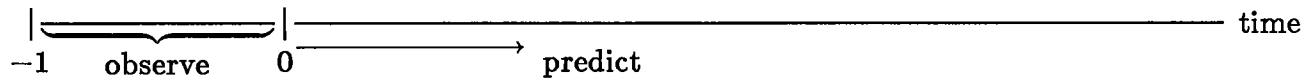
Validation of predictors on subsequent years data

year	# times LS better	# times EB better	Ties
1973	32 (39%)	49 (60%)	1
1974	25 (30%)	57 (70%)	0

3. Efron - Thisted (1976)

In this application, the number of unseen species is modeled using EB techniques. Efron and Thisted give both NPEB and PEB estimates because, surprisingly, in their model the NPEB estimates can be explicitly calculated. Much of the paper is devoted to finding nonparametric estimates and assessing accuracy, uses techniques such as Euler's transformation and Linear Programming.

Model: Observe species for a specified time, then predict total.



$X_i(t), i = 1, \cdots, S$, is the number of times species i appears in the interval $[-1, t]$

$$x_i(t) \sim \text{ Poisson } [\lambda_i(1+t)]$$

$$\lambda_i \sim G(\lambda)$$

This model says that the number of observed species $X_i(t)$, is a function of both the length of time observed (t), and a species parameter, λ_i. The model also specifies a common distribution of λ_i, allowing us to extrapolate to unseen species.

NPEB Approach

Define n_x = number of species observed exactly x times in $[-1, 0]$. Then it is straightforward to compute the expected value of n_x, η_x, as

$$\eta_x = E(n_x) = S \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} dG(\lambda),$$

where the form of $G(\lambda)$ is unspecified. The parameter of interest, however, is not η_x, but rather $\Delta(t)$, where

$$\begin{aligned}\Delta(t) = &\text{ expected number of species observed in } (0,t] \textit{ but not in } [-1,0]\\ = &\text{ expected number of new species in next } t \text{ time units}\\ = &\, S\int_0^\infty e^{-\lambda}(1-e^{-\lambda t})dG(\lambda)\end{aligned}$$

Efron and Thisted get NPEB estimate without specifying G. By expanding $(1-e^{-\lambda t})$ in a Taylor series and, using the previous expression, for η_x, we get

$$\Delta(t) = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 - + \cdots$$

suggesting the estimator

$$\hat{\Delta}(t) = n_1 t - n_2 t^2 + n_3 t^3 - + \cdots$$

As a constrast, in the PEB approach we specify a form for $G(\lambda)$. A convenient choice is $G(\lambda) \sim$ Gamma (α, β). This yields

$$\begin{aligned}\eta_x &= \frac{\eta_1 \Gamma(x+\alpha)}{x!\Gamma(1-\alpha)}\gamma^{x-1}, \qquad \gamma = \frac{\beta}{1+\beta}\\ \Delta(t) &= \frac{n_1}{\gamma\alpha}[(1+\gamma t)^{-\alpha} - 1]\end{aligned}$$

Continuing in a standard EB way, the parameters α and $\gamma = \frac{\beta}{1+\beta}$ are estimated from the marginal distribution of the x's, which is negative binomial. Efron and Thisted use maximum likelihood to do this.

The data that Efron and Thisted use is quite interesting. They equate unseen species = words that Shakespeare knew but did not use. Therefore, $\Delta(1)$ = expected number of new words that would be found in a volume of Shakespeare equal in size to his known work.

The data, $n_x \quad x = 1, \cdots, 100$, are available, and using these data we get

NPEB estimate: $\hat{\Delta}(1) = 11,430$

PEB estimate: $\hat{\Delta}(1) = 11,483$

Note that the PEB estimate only uses n_1 explicitly, that is,

$$\hat{\Delta}(1) = \frac{-n_1}{\hat{\gamma}\hat{\alpha}}[(1+\hat{\gamma})^{-\hat{\alpha}} - 1]$$

where $\hat{\alpha} = -.3954$, $\hat{\gamma} = .9950$ from ML (using all data).

The predictors from the PEB estimator are remarkably good, and are summarized below.

$x = \#$ times species (word) observed

Frequency	1	2	3	4	5	6	7
Observed	14376	4305	2281	1471	1050	798	633
Predicted	14399	4343	2292	1463	1043	837	638

VI. Empirical Bayes and the Stein Effect

The Stein Effect is a phenomenon that shows that, in certain cases, estimates can be improved by combining problems, and is the theoretical basis for the optimality of EB. Suppose that we have observations from p independent problems

$$\left.\begin{array}{l} x_1 \sim n(\theta_1, 1) \\ x_2 \sim n(\theta_2, 1) \\ \quad \vdots \\ x_p \sim n(\theta_p, 1) \end{array}\right\} \text{ independent normal populations.}$$

The estimation problem is to estimate $\theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}$ with an estimator $\delta = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_p \end{pmatrix}$, where the worth of the estimator δ is measured by a loss function

$$\sum_{i=1}^{p} (\delta_i - \theta_i)^2.$$

Associated with this loss is a risk function, the expected value of the loss

$$\text{Risk } = E_\theta \left[\sum_{i=1}^{p} (\delta_i - \theta_i)^2 \right]$$

The usual estimate of θ is $\delta^0 = X = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$, the estimate that corresponds, for example, to using the cell means to estimate treatment means in an ANOVA. For this estimator we have

$$\text{Risk of } X = E_\theta \left[\sum_{i=1}^{p} (x_i - \theta_i)^2 \right] = p$$

The estimator X possesses many optimality properties, but Stein's estimator is closer, on the average. The earliest, and simplest version of the estimator is to estimate θ_j with

$$\delta_j^s = \left(1 - \frac{p-2}{\sum_{i=1}^{p} x_i^2}\right) x_j \quad \longleftarrow \quad \text{from } j^{th} \text{ problem}$$

combines
all problems

This estimator satisfies

$$E_\theta \left[\sum_{i=1}^{p} (\delta_i^s - \theta_i)^2\right] < p \quad \text{for all } \theta,$$

so it is uniformly better than X.

The calculations for the proof are easiest in the normal case, but the "Stein Phenomenon" is quite extensive, covering many distributions and many loss functions. In particular, the following cases have been treated (see the references).

Poisson: Peng, Clevinson-Zidek

Neg. Binomial: Tsui

Exponential Families: Hudson, Berger

Discrete Exponential Families: Hwang

Gamma: Berger

Spherically Symmetric Distributions: Brandwein-Strawderman

Convex Loss: Brandwein-Strawderman

Absolute Error Loss: Berger

The EB explanation of Stein's Estimator gives us some insight into why the estimator is an improvement. Consider the following sample case.

Model

$$x_i \sim n(\theta_i, 1) \qquad i = 1, \ldots, p$$

$$\theta_i \sim n(0, \tau^2)$$

where the single distribution for the θ_i serves to the problems together. The posterior distribution $\theta_i|x_i$ is normal with mean

$$E(\theta_i|x_i,\tau^2) = \frac{\tau^2}{\tau^2+1}x_i = \left(1 - \frac{1}{\tau^2+1}\right)x_i.$$

The marginal distribution of x_i is normal,

$$x_i \sim n(\theta, \tau^2+1), \text{ independent}$$

and so it follows that $\sum x_i^2$ has a chi squared, distribution. That is $\sum_{i=1}^{p} x_i^2 \sim (\tau^2+1)\chi_p^2$, and a calculation shows

$$E\left(\frac{p-2}{\sum_{i=1}^{p} x_i^2}\right) = \frac{1}{\tau^2+1}.$$

Substituting this estimate into $E(\theta_i|x_i,\tau^2)$ yields

$$E(\theta_i|x_i,\hat{\tau}^2) = \left(1 - \frac{p-2}{\sum_{i=1}^{p} x_i^2}\right)x_i, \text{ Stein's Estimator}$$

The typical Risk behavior of Stein's estimator is pictured below in Figure 6.1.

FIGURE 6.1 ABOUT HERE

We can see that there is good risk improvement if the prior information correct, that is, if θ is near zero. However, we cannot do worse than usual estimate, since the risk of Stein's estimator is always below that of X. This is a real advantage. We are not severely penalized for a wrong prior guess.

A problem with the early use was that people didn't realize that choice of place to shrink (submodel) was very important. The region of risk improvement is quite narrow — thus if the submodel is wrong, the EB estimator quickly collapses back to usual estimator.

VII. A General EB ANOVA (Regression) Model

We now look at a more general model of the form

$$Y_i \sim \text{ independent } n(\theta_i, \sigma^2/n_i) \qquad i = 1, \ldots, p$$

($Y_i =$ observed ANOVA cell means, $\theta_i =$ true means,

$n_i =$ # observations/cell).

For now, assume σ^2 known, $n_i = 1$. Calculations can be done in a more general case (see, e.g. Lindley and Smith, 1972 *JRSSB* or, DuMouchel and Harris, 1983 *JASA*) given by

$$\begin{aligned} y_i &= \theta_i + \varepsilon_i \quad i = i, \ldots, p \\ \theta_i &= z_i'\beta + \delta_i \quad i = 1, \ldots, p \end{aligned}$$

↑— more flexible submodel for θ 's.

We would like to have the dimension of β be as small as possible to obtain greatest improvement in risk. Also, we want the submodel to have a chance of being true, which is accomplished by increasing the dimension of β. Thus, we have opposing goals.

The model can be generalized to the form

$$Y_i = X_i'\theta_i + \varepsilon_i$$

$$\theta_i = z_i'\beta + \delta_i$$

with only an increase in algebraic effort.

Common distributional assumptions are

$$Y_i|\theta_i \sim n(\theta_i, \sigma^2),$$

$$\theta_i|\beta \sim n(z_i'\beta, \tau^2),$$

$$\beta_i \sim \text{ uniform } (-\infty, \infty),$$

where $z_i = r \times 1$ vector of known predictor variables, $\tau^2 =$ unknown, $\beta = r \times 1$ vector of unknown regression coefficients. Using matrix notation we can write

$$Y|\theta \sim N(\theta, \sigma^2 I)$$

$$\theta|\beta \sim N(z'\beta, \tau^2 I)$$

$$\beta \sim \text{ uniform } \mathbb{R}^r$$

Note: Again, with an increased amount of algebra one can have an even more complicated covariance structure.

Now

$$E(\theta|Y) = \text{ formal Bayes estimator of } \theta, \text{ the vector of means}$$
$$= \int \underbrace{\theta\pi(\theta|Y)}\, d\theta$$

where

$$\pi(\theta|Y) = \frac{f(Y,\theta)}{m(Y)} = \frac{f(Y|\theta)\pi(\theta)}{m(Y)} \quad \text{(posterior distribution of } \theta)$$
$$f(Y|\theta) = \text{ sampling distribution}$$
$$\pi(\theta) = \text{ prior distribution } = \int \pi(\theta|\beta)d\beta$$
$$m(Y) = \text{ marginal distribution } = \int_\theta f(Y,\theta)d\theta.$$

A lengthy calculation shows that $\pi(\theta|Y)$ is normal with

$$\text{mean } = \frac{\tau^2}{\sigma^2+\tau^2}\left(I + \frac{\sigma^2}{\tau^2}H\right)Y$$
$$\text{covariance matrix } = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\left(I - \frac{\sigma^2}{\sigma^2+\tau^2}H\right)^{-1}$$

where

$$H = Z(Z'Z)^{-1}Z'.$$

Thus, the Bayes estimator of Y, the posterior mean, is

$$E(\theta|Y) = \frac{\tau^2}{\sigma^2+\tau^2}\left(I + \frac{\sigma^2}{\tau^2}H\right)Y = HY + \left(1 - \frac{\sigma^2}{\sigma^2+\tau^2}\right)(Y - HY).$$

(Note that the quantity HY is easily obtained by regressing Y on Z.) The EB estimate is obtained by replacing the unknown quantity $\left(1 - \frac{\sigma^2}{\sigma^2+\tau^2}\right)$ by an estimate based on the marginal (unconditional on θ) distribution of Y. Marginally, Y is singular normal, but the quadratic form

$$\frac{Y'(I-H)Y}{\sigma^2+\tau^2} \sim \chi^2_{p-r},$$

and a standard calculation shows

$$E\left[\frac{(p-r-2)\sigma^2}{Y'(I-H)Y}\right] = \frac{\sigma^2}{\sigma^2+\tau^2},$$

so we have an unbiased estimator of $\sigma^2/(\sigma^2+\tau^2)$.

An empirical Bayes estimator of θ can be constructed as

$$E(\theta|Y,\hat{\tau}) = HY + \left(1 - \frac{(p-r-2)\sigma^2}{Y'(I-H)Y}\right)(Y-HY).$$

The modification

$$HY + \left(1 - \frac{(p-r-2)\sigma^2}{Y'(I-H)Y}\right)^+ (Y-HY),$$

where $(x)^+ =$ maximum $(0,x)$, gives a uniform improvement in risk performance.

For unknown σ^2, and $n_i \neq 1$, the theory remains essentially unchanged, except that the algebra gets more difficult. Using a similar development, an empirical Bayes estimator is

$$H_0Y + \left(1 - \frac{\frac{\nu}{\nu+2}(p-r-2)s^2}{Y'(I-H_0)'D(I-H_0)Y}\right)^+ (Y-H_0Y)$$

where

$$\begin{aligned} D &= \text{ diagonal } (n_1, n_2, \ldots, n_p) \\ H_0 &= Z(Z'DZ)^{-1}Z'D \\ s^2 &= \text{ unbiased estimate of } \sigma^2 \text{ from full model} \\ \nu &= \text{ error df.} \end{aligned}$$

Example: Growth Curves (Strenio, et al., 1983 *Biometrics*)

Model: Growth is modeled for each individual as a polynomial in age. A second-stage model relates individuals using covariates. Strenio, et al. do calculations in general, using some of the Lindley-Smith calculations. We will just do an example.

Let Y_{it} = weight of i^{th} rat at t^{th} week. Then we have

$$\begin{aligned} E(Y_{it}|\pi_{1i},\pi_{2i}) &= \pi_{1i} + \pi_{2i}(t-1) \\ E(\pi_{1i}|\gamma_{11},\gamma_{12}) &= \gamma_{11} + \gamma_{22}x_i \\ E(\pi_{2i}|\gamma_{21},\gamma_{22}) &= \gamma_{21} + \gamma_{22}x_i \\ x_i &= \text{ mother's weight (covariate)} \end{aligned}$$

The EB estimate of π_{1i} is a linear combination of

$$\hat{\pi}_{1i} \text{ (from regression of } i^{th} \text{ rat on time)}$$

and

$$\hat{\gamma}_{11} + \hat{\gamma}_{12} x_i \text{ (from regression of } \hat{\gamma}_{1i} \text{ on } x_i).$$

The EB estimate of π_{2i} similarly obtained. The growth curve (a line in this case) is thus a linear combination of the curve for the individual and the curve for the ensemble. Note that the submodel estimates. $\hat{\gamma}_{11}$ and $\hat{\gamma}_{12}$, are based on data that is summed over time, hence use all of the information.

FIGURE 7.1 ABOUT HERE

VIII. Estimates of Variance

A conservative approach to attaching variance estimates to EB estimates is the following. Many EB (Stein-type) estimators dominate the usual estimators in mean squared error (MSE) i.e.

$$\text{MSE(EB estimator)} < \text{MSE(Usual estimator)}.$$

Since

$$\text{MSE} = \text{Variance} + (\text{Bias})^2,$$

EB estimates, which are biased, have smaller variance. Therefore, by using the usual estimates of variance one is being conservative (i.e., the estimate may be an overestimate, but not an underestimate). This means that one can just "recenter" the usual interval at an EB estimate and obtain an interval estimate with higher coverage probability.

While the above approach is reasonable in theory, in practice we would like more. For example, it should be possible to produce smaller variance estimates, and hence shorter confidence intervals.

We can dominate the usual <u>simultaneous</u> (Scheffé) procedure by again taking advantage of the Stein-Effect. However, we <u>cannot</u> dominate componentwise — the usual one-dimensional confidence interval is admissible (cannot be uniformly dominated).

Componentwise Intervals have been derived by Morris in the following way.

For the EB Anova Model,

$$Y_i|\theta_i \sim n(\theta_i, \sigma^2/n) \qquad i = 1, \ldots, p$$
$$\theta_i|\beta \sim n(z_1'\beta, \tau^2)$$
$$\beta \sim \text{ uniform}$$

where Y_i = cell means, n obs/cell, and

$$\beta = r \times 1 (r < p) \text{ vector of unknown regression coefficients,}$$

the Bayes estimator is

$$\hat{\theta}_{B_i} = \hat{\theta}_i + (1 - B)(Y_i - \hat{\theta}_i), \qquad B = \sigma^2/(\sigma^2 + n\tau^2)$$

where

$$\hat{\theta}_i = z_i'\hat{\beta}, \qquad \hat{\beta} = (z'z)^{-1}z'Y$$

The EB estimate of B, using similar arguments to the previous ones, is

$$\hat{B} = (p - r - 2)\left(\frac{\nu}{\nu + 2}\right)\frac{\hat{\sigma}^2}{n}/\sum(Y_i - \hat{\theta}_i)^2, \quad \nu = df \text{ for error,}$$

resulting in the EB estimator of θ_i,

$$\hat{\theta}_{EB_i} = \hat{\theta}_i + (1 - \hat{B})^+(Y_i - \hat{\theta}_i).$$

For the Bayes estimator, $\hat{\theta}_{B_i}$,

$$\text{Var}(\hat{\theta}_{B_i}) = \frac{\sigma^2}{n}(1 - B).$$

Morris (1983) notes that, for the EB estimator, we must also account for increase in variance due to

i) estimating B

ii) estimating β

and suggests

$$\text{Variance } (\hat{\theta}_{EB_i}) = \frac{\sigma^2}{n}\left(1 - \frac{p - r}{p}\hat{B}\right) + v(\hat{B})(Y_i - \hat{\theta}_i)^2$$

$$\text{with } v(\hat{B}) = \frac{2}{p-r-2}\hat{B}^2$$

Note: 1. Even in the equal variance model, the EB variance estimates may be unequal.

2. In some coordinates the EB variance estimate will be smaller than the usual estimate, in others it will be larger.

For the steer data, $\hat{\sigma}/\sqrt{n} = 1.24$. The EB standard deviations are:

	Treatment				
	1	2	3	4	5
EB St. Dev.	1.23	1.32	1.13	1.13	1.14

↑ This cell had the mean that was furthest from the linear submodel.

We can also use EB methods to improve on the usual methods for simultaneous intervals in ANOVA. Higher dimensional (≥ 3) EB confidence sets can be constructed that dominate the usual Scheffé procedure. Such results again take advantage of the Stein effect. (Hwang and Casella, 1987)

In particular, recentering a confidence sphere at an EB (Stein) estimator results in a uniform increase in coverage probability. Reduction of the radius of a confidence set, while maintaining improved coverage probability, is also possible (Casella and Hwang, 1983).

A confidence set on a vector of means is equivalent to a set of simultaneous intervals. In particular, the Scheffé method of constructing simultaneous intervals yields:

$$\theta_i \in Y_i \pm s\sqrt{pF_\alpha} \quad i = 1, \ldots, p$$

with probability $1-\alpha$ simultaneously. Here F_α is the upper α cut off from an F distribution with p and ν degrees of freedom.

Using EB methodology, we can construct simultaneous intervals

$$\theta_i \in \hat{\theta}_{EBi} \pm sV(Y, s), \quad i = 1, \ldots, p$$

with probability $1-\alpha$ simultaneously, where $V(Y,S) \leq \sqrt{pF\alpha}$. We can also obtain intervals on contrasts, for example, pairwise differences numbers for steer data presented earlier.

Casella and Hwang (1983), derive an EB confidence function given by

$$V_{EB}^2 = \left(1 - \frac{a}{pF\alpha}\right)\left[pF\alpha - p\log\left(1 - \frac{a}{pF\alpha}\right)\right] \quad \text{if } T \leq pF\alpha$$

$$= \left(1 - \frac{a}{T}\right)\left[pF\alpha - p\log\left(1 - \frac{a}{T}\right)\right] \quad \text{if } T > pF\alpha$$

where, $T = \sum(Y_i - \hat{\theta}_i)^2/S^2$, $a = \frac{\nu}{\nu+2}(p-2)$. Using this confidence function, intervals of the form

$$\theta_i \in \hat{\theta}_{EBi} \pm sV_{EB}(T)$$

can be constructed. These intervals maintain $1-\alpha$ confidence while providing a reduction in length.

In general, confidence statements for EB procedures are in a primitive state. One reason for this is that the small sample distributions are quite difficult to deal with and theoretical properties are hard to verify. This difficulty is gradually being overcome, however, both with improved analytic techniques and careful use of the computer. Many small sample properties of EB estimates have been investigated by careful computer studies.

The other reason for the slow development of EB confidence statements, perhaps the more serious problem, has to do with the meaning of the word confidence. There is no overall agreement as to whether EB confidence sets should provide frequency confidence (long-run guarantees), Bayesian confidence (conditional guarantees, given the observed data), or some mixture of the two. At present, there is no clear solution to this problem, a problem that has implications in the foundations of statistics.

There is probably no one correct answer to this problem. EB models are extremely helpful in obtaining good estimators in complicated situations. The ultimate inference, be it frequency, Bayesian, or some other, can be a matter of choice.

IX. Other Applications

1. Contingency Tables. The usual contingency table can be described by

$$x_{ij} = \text{ observed frequency in cell } (i,j),$$

where x_{ij} is multinomial with $Ex_{ij} = Np_{ij}$, $\sum_{i,j} x_{ij} = N$, $\sum_{i,j} p_{ij} = 1$

A log-linear model for the cell probabilities is

$$\log p_{ij} = u_0 + \underset{\substack{\uparrow \\ \text{row} \\ \text{effect}}}{u_{1i}} + \underset{\substack{\uparrow \\ \text{column} \\ \text{effect}}}{u_{2j}} + \underset{\substack{\uparrow \\ \text{interaction} \\ \text{effect}}}{u_{12ij}}$$

EB methods for contingency tables now place prior distributions on the u's and estimate unknown prior parameters from the marginal distribution of the x_{ij}'s.

A popular method (Bishop, Feinberg and Holland) is to use a Dirichlet prior. This can yield estimates of the form

$$\hat{n}_{EB} = \hat{\omega}_{ij} + (1 - \hat{\omega})\hat{n}_m,$$

where $\hat{n}_m$ is the maximum likelihood estimator under independence, and $\hat{\omega}$ is estimated from the marginal distribution. Such methods are particularly helpful with large, sparse tables.

Leonard (1975) uses a Bayesian approach with prior distributions

$$u_{1i} \sim n(\mu_1, \sigma_1^2),$$
$$u_{2j} \sim n(\mu_2, \sigma_2^2),$$
$$u_{12ij} \sim N(\mu_3, \sigma_3^2),$$

where, μ_i uniform $(-\infty, \infty)$, $\sigma_i^2 \sim$ inverse χ^2.

Laird (1978) uses an EB approach with

$$u_{1i} \sim \text{ Uniform } (-\infty, \infty),$$
$$u_{2j} \sim \text{ Uniform } (-\infty, \infty),$$
$$u_{12ij} \sim n(0, \sigma^2),$$

and σ^2 is then estimated from the marginal distribution. Much computation is involved in getting estimates.

2. ANOVA: The case of unequal variances. The usual ANOVA model, with unequal variances, can be described by

$$\text{Model: } y_{ij} = \theta_i + \varepsilon_{ij} \qquad \begin{matrix} i = 1, \ldots, p \\ j = 1, \ldots, n_i \end{matrix}$$

$y_{ij} = jth$ observation on ith treatment

$\theta_i = ith$ treatment mean

$\varepsilon_{ij} \sim n(0, \sigma_i^2)$

↑ variance depends on treatment

If $\sigma_i^2 = \sigma^2$, we have an unbalanced ANOVA, since $n_i \neq n$. Since the observations having common variances, the previously mentioned methodology applies. In fact, an EB estimator for this case has already been presented. If we assume, however, that the σ_i^2 are totally unknown we run into additional complications.

With the above model, make the Bayesian assumption

$$\theta_i \sim n(\mu, \tau^2).$$

It then follows that the Bayes estimator of θ_i is

$$\hat{\theta}_{B_i} = \mu + \Big(1 - \frac{\sigma_i^2/n_i}{\sigma_i^2/n_i + \tau^2}\Big)(\overline{y}_i - \mu)$$

and marginally,

$$\overline{y}_i \sim n\Big(\mu,\ \frac{\sigma_i^2}{n_i} + \tau^2\Big)$$

The problem now is that the $\overline{y}_i$ are not marginally identically distributed, so we cannot get an estimate for $\frac{\sigma_i^2}{n_i} + \tau^2$, based on all the data, as easily as in the equal variance case.

Let $S_i^2 =$ Sum of squares in ith treatment,

$$S_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2.$$

Marginally,

$$S_i^2 \sim (\sigma_i^2 + \tau^2)\chi^2_{n_{i-1}}$$

and we use the joint (marginal) distribution of $S_1^2, \ldots, S_p^2$ to obtain a maximum likelihood estimate of τ^2.

One can now estimate $\sigma_i^2/(\sigma_i^2 + n_i\tau^2)$ by

$$\frac{(p-2)\sigma_i^2}{\sigma_i^2 + n_i\tilde{\tau}^2} \text{ if } \sigma_i^2 \text{ are known}$$

$$\frac{\nu_i}{\nu_{i+2}} \frac{(p-2)\hat{\sigma}_i^2}{(\hat{\sigma}_i^2 + n_i\tilde{\tau}^2)} \text{ if } \sigma_i^2 \text{ are unknown, } \nu_i = n_i - 1$$

Morris (1983) gives generalizations of this. Rubin (1980) also considers this case, and uses the EM algorithm to obtain estimates.

3. Example: Toxoplasmosis prevalence rates (Efron-Morris, 1975). Toxoplasmosis rates were estimated in 36 cities in El Salvador, based on sample of 5171 individuals.

Data are prevalence rates adjusted for age distribution in each city, and are binomial. Also, we have unequal variances result since there are

a) different prevalence rates in each city

and

b) different sample size in each city.

Let X_i = adjusted prevalence rate. Assume

$$X_i|\theta_i \sim n(\theta_i, \sigma_i^2)$$

$$\theta_i \sim N(0, \tau^2)$$

EB estimate is then

$$\frac{\sigma_i^2}{\sigma_i^2 + \hat{\tau}^2} X_i$$

$\uparrow$ marginal MLE

FIGURE 9.1 ABOUT HERE

FIGURE 9.2 ABOUT HERE

Selected Estimates and
Empirical Bayes Estimates of
Toxoplasmosis Prevalence Rates

City	X_i	σ_i	EB_i
1	.293	.304	.035
4	.152	.115	.075
5	.139	.081	.092
8	.098	.087	.062
13	.035	.056	.028
21	-.034	.073	-.024
25	-.098	.068	-.072
28	-.138	.063	-.106
29	-.156	.077	-.107
31	-.241	.106	-.128
32	-.294	.179	-.083
33	-.296	.064	-.225

Notice how the maximum likelihood estimates with larger variance get shrunk more – in particular note the change in estimates for cities 1 and 32. City 33 has virtually the same ML estimate as city 32, but a much smaller variance. Hence it has shrunk very little. If we had assumed $\sigma_i^2 = \sigma^2$, each estimate (x_i) would have been shrunk by an equal, smaller amount.

4. Example: Estimation of Failure Times (Basu-Rigdon). We have N systems, and we want to estimate failure rates of these systems. Let $\lambda_1, \ldots, \lambda_N$ denote the failure rates (mean # of failures). For each system observe failure times

$$t_1^i, \ldots, t_{n_i}^i \quad i = 1, \ldots, N$$

(observe until n_i failures occur)

Model:

$$f(t_1^i, \ldots, t_{n_i}^i | \lambda_i) = \lambda_i^{n_i} e^{-\lambda_i t_{n_i}^i} \qquad i = 1, \ldots, N$$
$$\pi(\lambda_i | \alpha, \theta) = \frac{\theta^\alpha}{\Gamma(\alpha)} \lambda_i^{\alpha-1} e^{-\theta\lambda_i}$$

The estimates of θ and α obtained from marginal distribution, and the authors use a Newton-Raphson algorithm to obtain ML estimates of θ and α. Under the Bayes model, the posterior mean (Bayes estimate) is

$$\begin{aligned} E(\lambda_i|t_{n_i}) &= \frac{n_i+\alpha}{t_{n_i}+\theta} \\ &= \left(\frac{t_{n_i}}{t_{n_i}+\theta}\right)\left(\frac{n_i}{t_{n_i}}\right) + \left(\frac{\theta}{t_{n_i}+\theta}\right)\left(\frac{\alpha}{\theta}\right) \end{aligned}$$

which is a weighted average of

$$\frac{n_i}{t_{n_i}} = \text{ MLE of } \lambda_i$$

and

$$\frac{\alpha}{\theta} = \text{ prior mean of } \lambda_i$$

Note that as t_{n_i} increases, the MLE is weighted more. The EB estimate is obtained by substituting estimate of α and θ into the above equation. Calculations are also done for log-normal prior distribution, where the EM algorithm is used to obtain marginal MLEs.

Selected Operating Times Between Failures or Airconditioning Equipment in Boeing 720 Aircraft

Plane Number						
1	3	5	7	9	11	13
194	90	55	97	359	130	102
15	10	320	51	9	493	209
41	60	56	11	12		14
29	186	104	4	270		57
33	61	220	141	603		54
181	49	239	18	3		32
	14	47	142	104		67
	24	246	68	2		59
	56	176	77	438		134
	20	182	80			152
	79	33	1			27
	84	15	16			14
	44	104	106			230
	59	35	206			66
	29		82			61
	118		54			34
	25		31			
	156		216			
	310		46			
	76		111			
	26		39			
	44		63			
	23		18			
	62		191			
	130		18			
	208		168			
	70		24			
	101					
	208					

Classical and PEB Point and Interval
Estimates of Failures per 1000 Hours
for Aircraft Airconditioning Data

Plane No.	Total Time	Classical MLE	Classical 95% C.I.	EB Pt.Est.	EB 95% P.I.
1	493	12.17	(4.46,23.67)	10.97	(7.06,15.72)
3	2422	11.97	(8.02,16.71)	11.41	(8.40,15.12)
5	1832	7.64	(4.18,12.13)	9.09	(6.23,12.48)
7	2074	13.02	(8.58,18.37)	11.93	(8.71,15.64)
9	1800	5.00	(2.29,8.76)	7.76	(5.13,10.92)
11	623	3.21	(0.39,8.94)	8.66	(5.32,12.81)
13	1312	12.20	(6.97,18.86)	11.30	(7.84,15.38)

5. Forestry. Predicting Hardwood Tree Volume from Diameter and Height (Green and Strawderman).

If a tree were a perfect cylinder

$$V = \frac{\pi}{4} D^2 H$$

would predict a volume perfectly. A widely used equation is

$$V = \beta_0 + \beta_1 (D^2 H)$$

Estimates of β_1 quite constant throughout literature so authors took $\beta_1 \sim$ normal (μ, σ^2) with μ known and equal to past averages. β_0 was given a uniform prior.

Model

$$V_i = \beta_{0i} + \beta_{1i}(D^2 H)_i + \varepsilon_i \qquad i = 1, \ldots, H = \# \text{ hardwoods}$$

$$\beta_{0i} \sim \text{ Uniform } (-\infty, \infty)$$

$$\beta_{1i} \sim \text{ Normal } (\mu, \sigma^2), \ \mu \text{ known}$$

Effectiveness of prediction equations was checked by "holding out" data. Specifically, the authors were interested in seeing if the EB estimates could predict as well as least squares but using fewer observations. Roughly speaking, EB estimates using 66% of the data were as good as least squares using all of the data.

X. Some References for EB Applications

Allen and Jordan (1982) *Biometrics*

A Bayesian approach to prediction, more precisely, extrapolation, in regression. Although no EB is done, the model used here can also be used to provide EB estimates. (normal)

Basu and Rigdon (Technical Report, Dept. of Stat., U. of Missouri, Columbia)

PEB techniques applied to failure time data. Two examples. (Poisson, gamma, log-normal)

Deely and Lindley (1981 *JASA*)

They argue that empirical Bayesians are really non-Bayesian, however, their arguments really apply to NPEB rather than PEB. They suggest some methodology remarkably similar to PEB, i.e., estimating from the marginal distribution. (normal, poisson, gamma)

DuMouchel and Harris (1983 *JASA*)

Apply Bayes and EB models to the problem of combining results from different cancer studies. Uses linear model-type theory. (normal)

Efron and Morris (1975 *JASA*)

Uses EB to justify Stein's estimator. Considers both equal and unequal variance cases. (normal)

Efron and Thisted (1976 *Biometrika*)

Uses both parametric and non-parametric EB models to estimate total # of words known to Shakespeare. Application to estimating total # of species based on counts of trapped species. PEB model attributed to Fisher. (poisson, gamma, negative binomial)

Green and Strawderman (1985 *Forest Science* 1986, *Canadian Journal of Forest Research*)

Applications of EB to estimate forest tree volume as a function of diameter and height. Use a mixture of EB and Bayes techniques. (normal)

Laird (1978 *Biometrika*)

EB methods in contingency tables. Applies normal and flat priors to log-linear model for 2-way tables, uses EM Algorithm to estimate from marginal distribution. (normal)

Rubin (1980 *JASA*)

EB prediction of first year Law School GPA. Good discussion of looking for a place to shrink toward. (normal)

Strenio, et al. (1983 *Biometrics*)

EB modeling of growth curves — combining curves for individual with group. Models a la Lindley and Smith (1972 *JRSSB*), but slightly more general. EM algorithm used for estimation of unknown covariance matrices. (normal)

REFERENCES

Allen, D. M. and Jordan D. C. (1982). The Use of Prior Information for Prediction, *Biometrics* **38**, p. 787.

Angers, J-F. (1987). Development of Robust Bayes Estimators for a Multivariate Normal Mean. Ph.D. Theses. Department of Statistics, Purdue University.

Basu, A. P., and Rigdon, S. E. (1985). Examples of Parametric Empirical Bayes Methods for the Estimation of Failure Processes for Repairable Systems. Technical Report, Dept. of Statistics, University of Missouri at Columbia.

Berger, J. O. (1976). Minimax Estimation of a Multivariate Normal Mean Under Arbitrary Quadratic Loss. *Journal of Multivariate Analysis* **6**, P. 256.

Berger, J. O. (1976). Tail Minimaxity in Location Vector Problems and its Applications. *Annals of Statistics* **4**, p. 33.

Berger, J. O. (1980). Improving on Inadmissible Estimators in Continuous Exponential Families with Applications to Simultaneous Estimation of Gamma Scale Parameters. *Annals of Statistics* **8**, p. 545.

Bishop, Y. M., Feinberg S. E., and Holland P. W. (1975). Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass: MIT Press.

Brandwein, A. and Strawderman, W. E. (1978). Minimax Estimation of Location Parameters for Spherically Symmetric Unimodal Distributions. *Annals of Statistics* **6**, p. 279.

Brandwein, A. and Strawderman, W. E. (1980). Minimax Estimation of Location Parameters for Spherically Symmetric Distributions with Concave Loss. *Annals of Statistics* **8**, p. 698.

Casella, G. and Hwang, J. T. (1983). Empirical Bayes Confidence Sets for the Mean of a Multivariate Normal Distribution. *Journal of the American Statistical Association* **78**, p. 688.

Casella, G., and Hwang, J. T. (1987). Employing Vague Prior Information in the Construction of Confidence Sets. *Journal of Multivariate Analysis* **21**, p. 79.

Clevenson, M. L. and Zidek, J. V. (1975). Simultaneous Estimation of the Means of Independent Poisson Laws. *Journal of the American Statistical Association* **70**, p. 698.

Deely, J. J. and Lindley D. V. (1981). Bayes Empirical Bayes. *Journal of the American Statistical Association* **76**, p. 833.

DuMouchel, W. M. and Harns, J. E. (1983). Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species (with discussion). *Journal of the American Statistical Associaton* **78**, p. 313.

Efron, B., and Morris, C. (1975). Data Analysis using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association* **70**, p. 311.

Efron, B., and Thisted, R. (1976). Estimating the Number of Unseen Species: How Many Words did Shakespeare Know? *Biometrika* **63**, p. 435.

Green, E., and Strawderman, W. E. (1985). The Use of Bayes Empirical Bayes in Individual Volume Equation Development *Forest Science* **31**, p. 975.

Green, E., and Strawderman, W. E. (1986). Stein Rule Estimation of Coefficients for Eighteen Eastern Hardwood Cubie Volume Equations. *Canadian Journal of Forest Research* **16**, p. 246.

Green, E., and Strawderman, W. E. (1986). Reducing Sample Size Through the Use of a Composite Estimator and Applications to Timber Volume Estimation. *Canadian Journal of Forest Research* **16**, p. 1116.

Hudson, H. M. (1978). A Natural Identity for Exponential Families with Applications in Multiparameter Estimator. *Annals of Statistics* **6**, p. 473.

Hwang, J. T. (1982). Improving Upon Standard Estimators in Discrete Exponential Families with Applications to Poisson and Negative Binomial Cases. *Annals of Statistics* **10**, p. 857.

Hwang, J. T. and Casella G. (1982). Minimax Confidence Sets for the Mean of a Multivariate Normal Distribution. *Annals of Statistics* **10**, p. 868.

Laird, N. M. (1978). Empirical Bayes Methods for Two-Way Contingency Tables. *Biometrika* **65**, p. 581.

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes Confidence Intervals Based on Bootstrap Samples (with discussion). *Journal of the American Statistical Association* **82**, p. 739.

Leonard, T. (1975). Bayesian Estimation Methods for Two-Way Contingency Tables.

Journal of the Royal Statistical Society, Series B **37**, p. 23.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, p. 1.

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association* **78**, p. 63.

Peng, J. C. M. (1975). Simultaneous Estimation of Parameters of Independent Poisson Distributions. Technical Report 78, Department of Statistics, Stanford University.

Rubin, D. B. (1980). Using Empirical Bayes Techniques in the Law School Validity Studies (with discussion). *Journal of the American Statistical Association* **75**, p. 801.

Strenio, J. F., Weisberg, H. I., and Bryk A. S. (1983). Empirical Bayes Etimation of Individual Growth-Curve Parameters and their Relationship to Covariates. *Biometrics* **39**, p. 71.

Tsui, K-W. (1984). Robustness of Clevenson-Zidek-Type Estimators. *Journal of the American Statistical Association* **79**, p. 152.

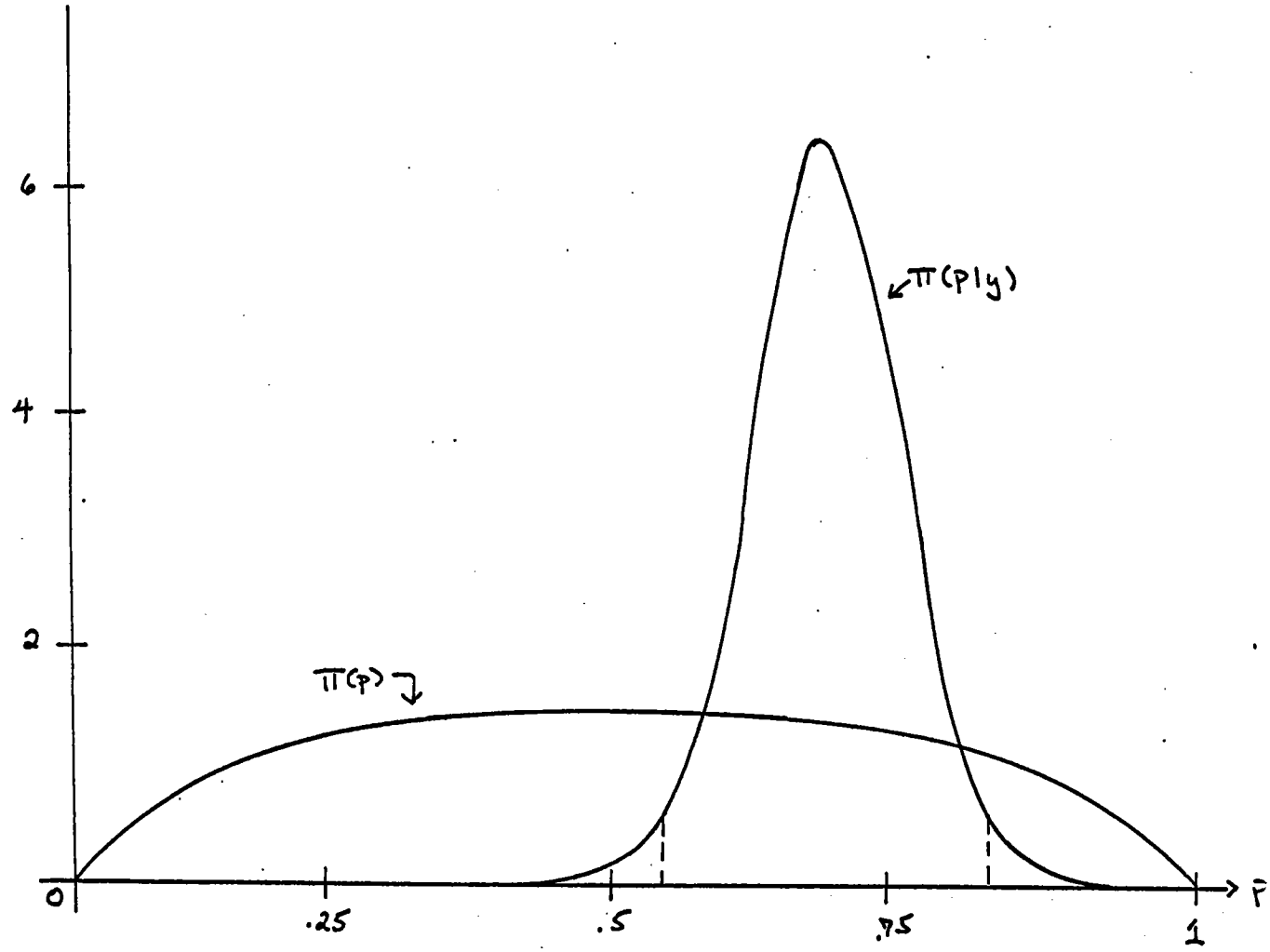


Figure 3.1: Bayes prior and posterior for binomial example.

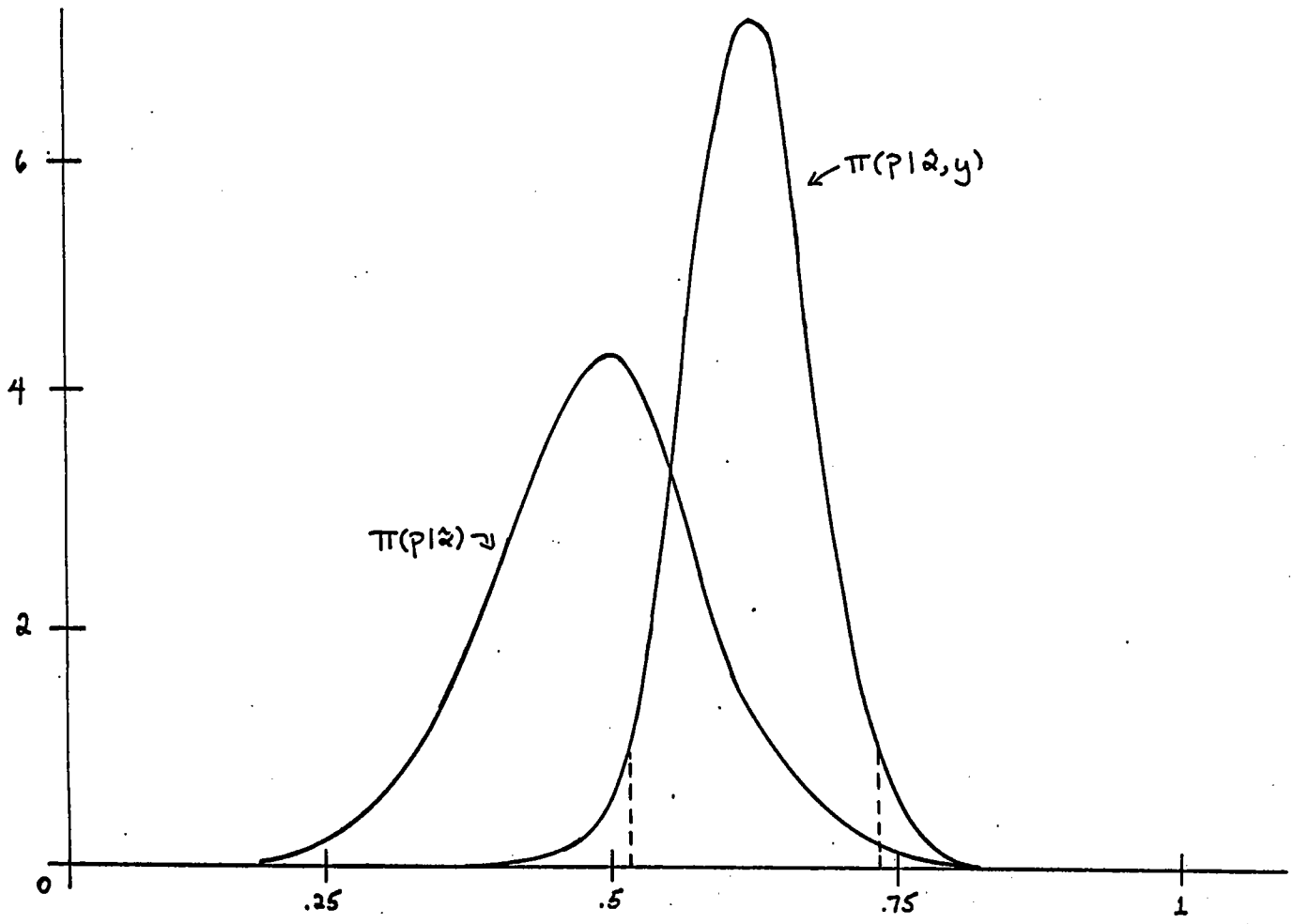


Figure 3.2: Empirical Bayes prior and posterior for binomial example.

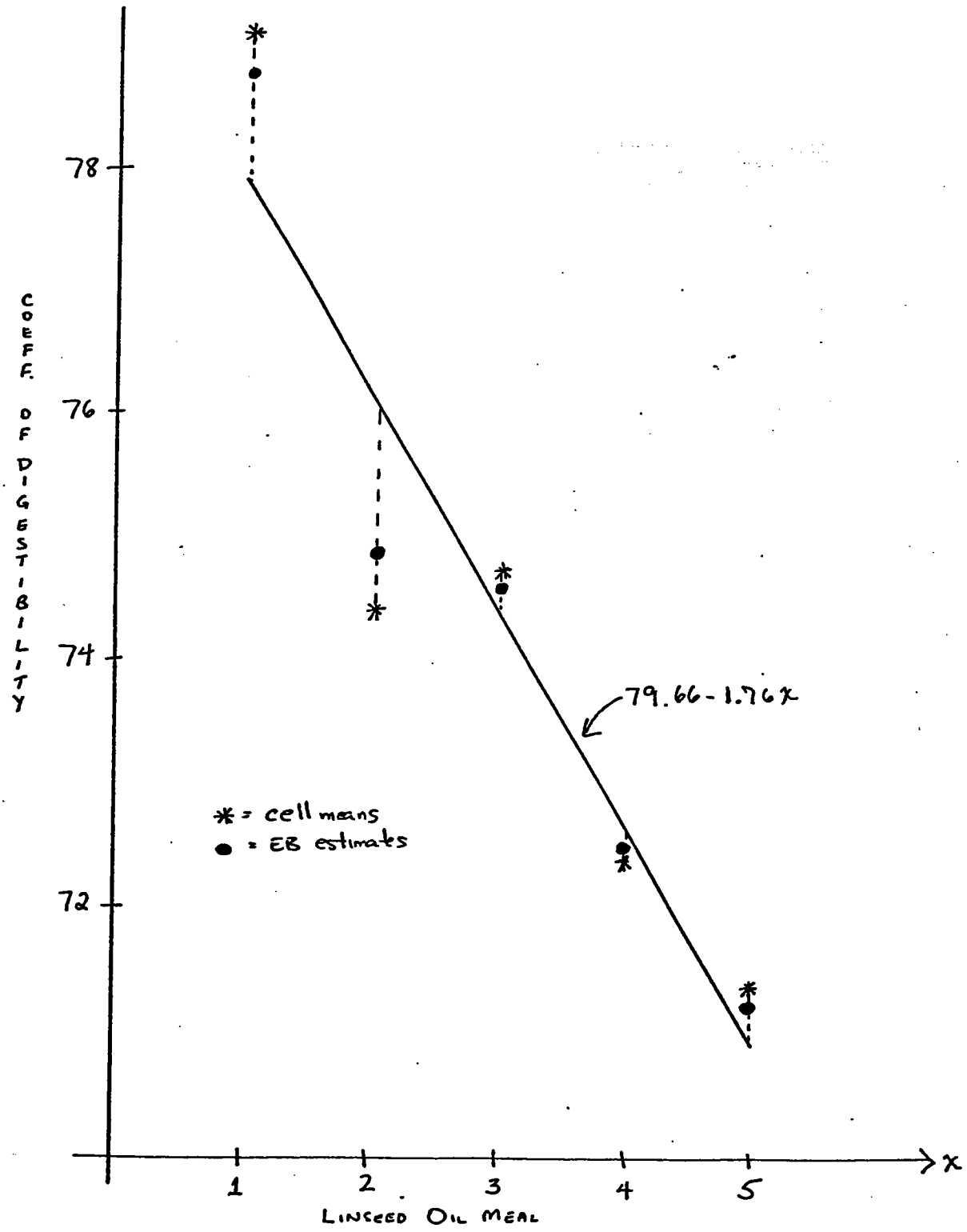


Figure 4.1: Estimates of the cell means for the steer data. The line represents the submodel.

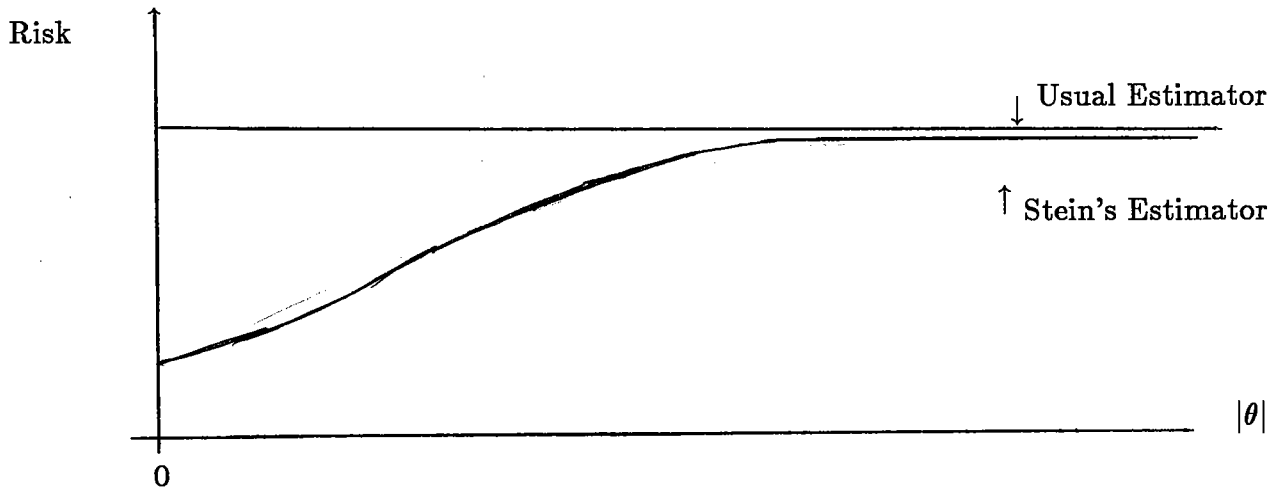


Figure 6.1: Risk functions of usual estimator and Steins estimator of a multivariate normal mean.

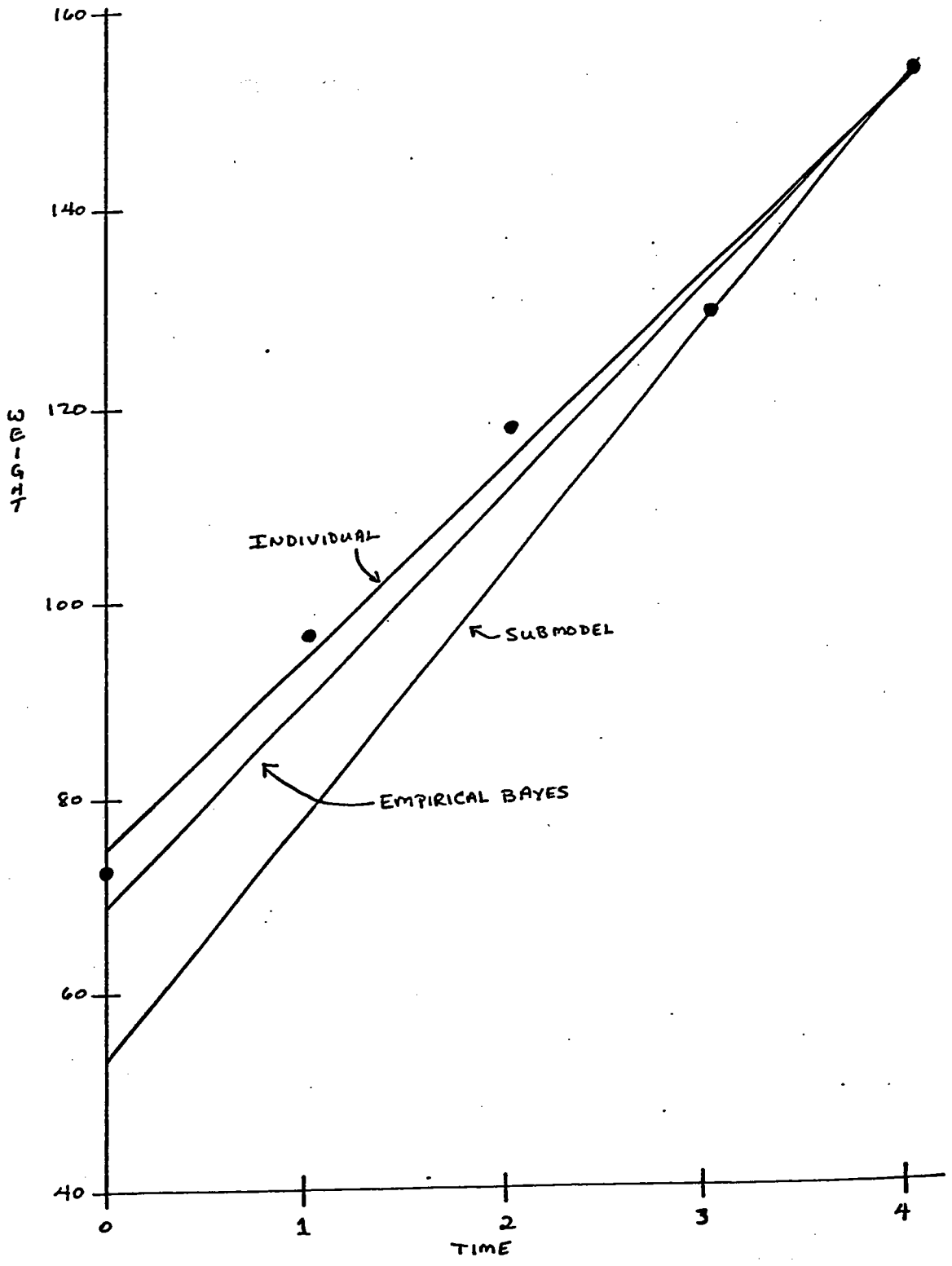


Figure 7.1: Typical growth curve for Strenio, et al. data. Lines pictured are for rat #8.

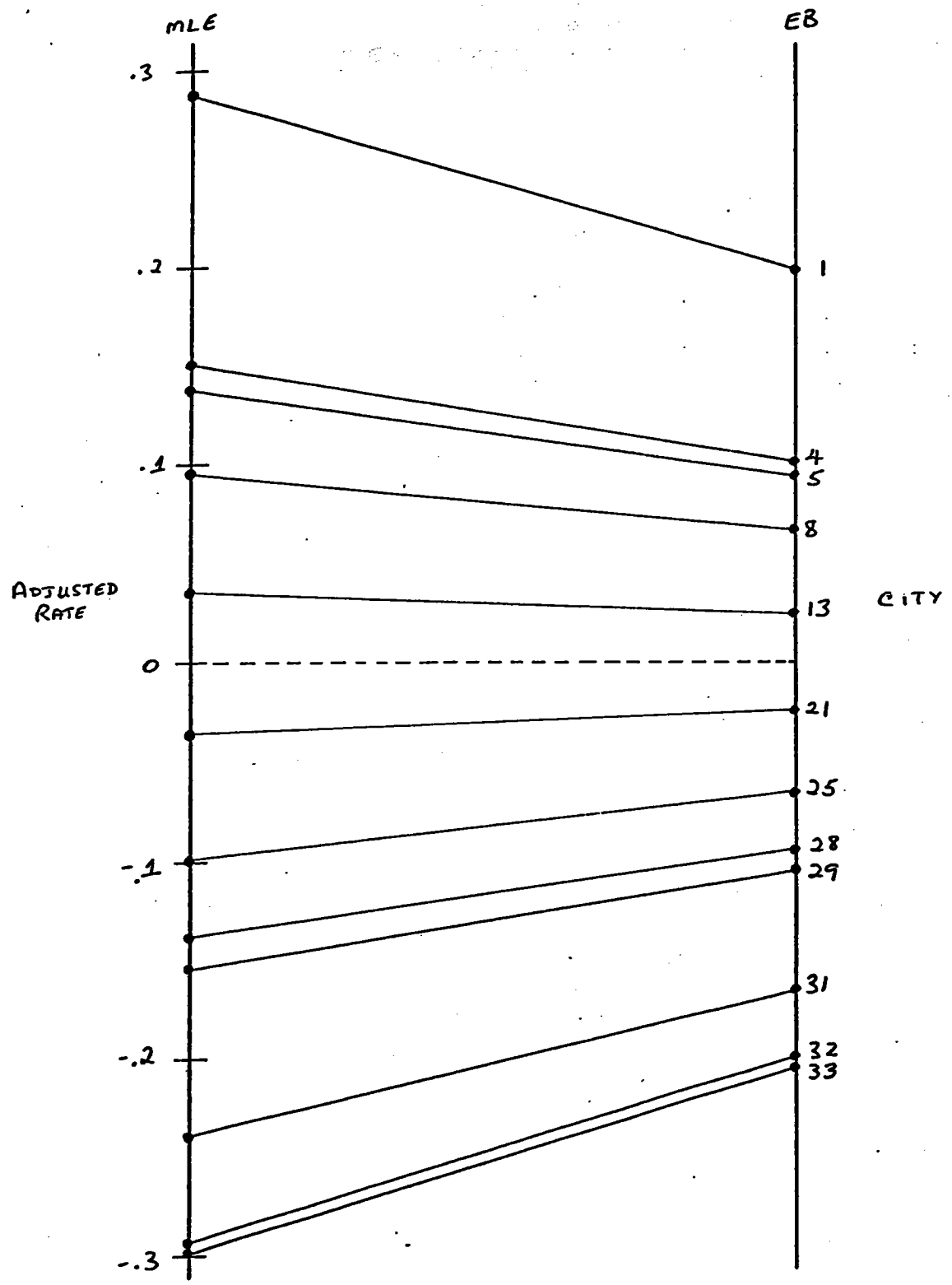


Figure 9.1: Estimated toxoplasmosis rates for selected cities — equal variance model.

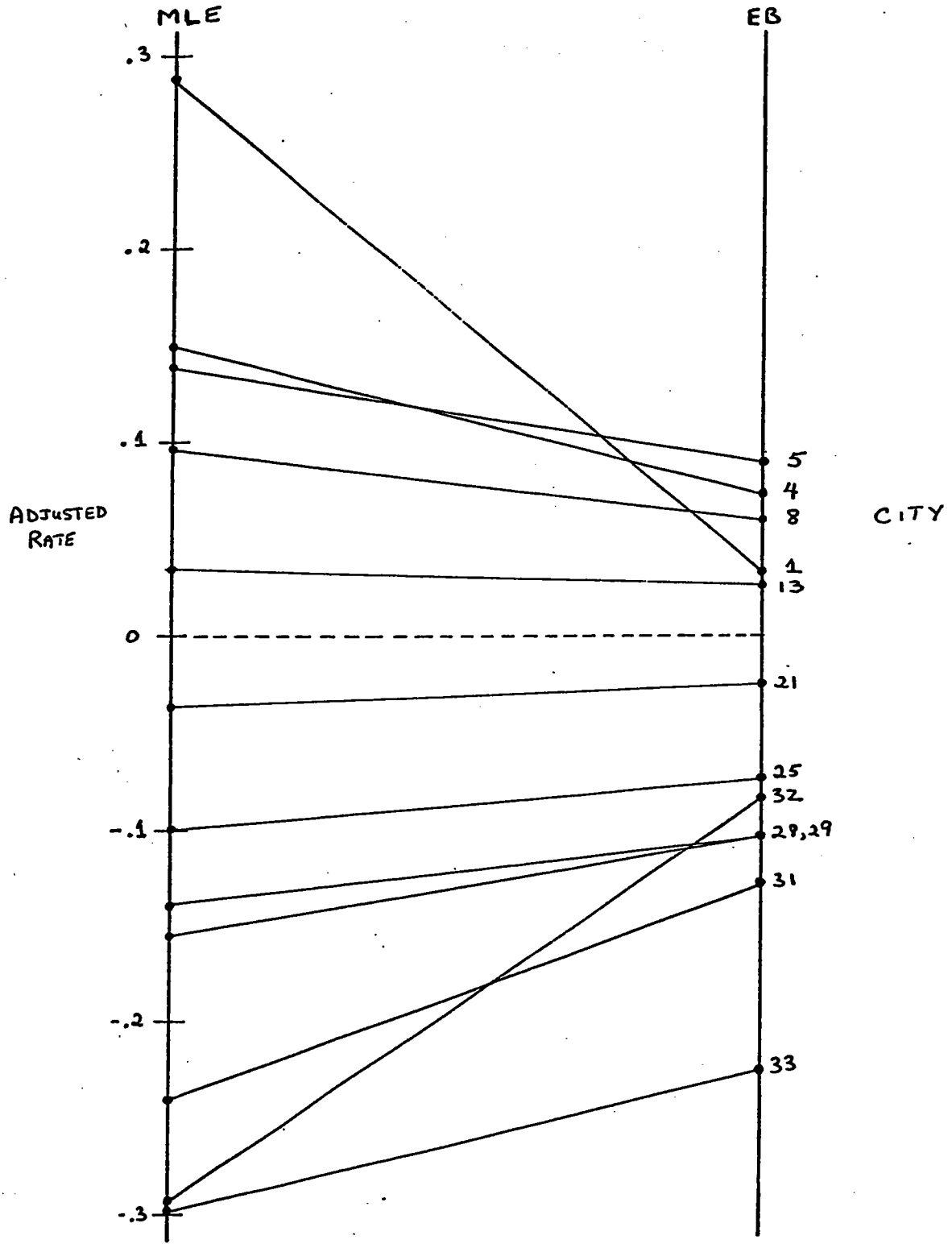


Figure 9.2: Estimated toxoplasmosis rates for selected cities — unequal variance model.