Suppose data arise via a ***two-stage sampling*** process:

$$[X \mid \eta] = \int [X \mid \theta][\theta \mid \eta] d\theta \,.$$

(Note the conditional independence factorization: $\eta \to \theta \to X$.)

Sometimes, your goal is to estimate η. Sometimes, your goal is to estimate θ rather than η.
The intermediate unknowns $\theta = (\theta_1,...,\theta_k)$ can represent missing data, incomplete (e.g. censored) data, latent (hidden) variables, random effects, …

Example η = prevalence of disease,
$\theta_1,\ldots,\theta_n$ =disease state of subjects (binary),
$X_1,\ldots, X_n$ = test results.

Three approaches

(1) η is known: standard Bayes
(2) η is unknown, but you have a prior for it: hierarchical Bayes
(3) η is unknown, but you will ***estimate*** it: empirical Bayes
(This is possible if θ and X are drawn more than once.)

Hierarchical Bayes

If η is "of interest", integrate out all the θ, values:

$$[\eta \mid X] = \int_\Theta [\theta,\eta \mid X] d\theta = [X \mid \eta][\eta] \,/\, [X] \quad \text{as usual.}$$

But if θ is of interest, then (of course) integrate out η instead:

$$[\theta \mid X] = \int_H [\theta,\eta \mid X] d\eta = \int_H [X \mid \theta][\theta \mid \eta][\eta] d\eta \,/\, [X] \;.$$

.

Empirical Bayes has two steps:

(E) "Empirical" Estimate (classically) η from the ***marginal*** likelihood:

$$[X \mid \eta] = \int [X \mid \theta][\theta \mid \eta] d\theta$$

Estimation methods:
Maximum likelihood.
Method of moments.

(B) "Bayes"

Do n "Bayes" analyses, pretending that you know η, standard frequentist "plug-in" style.

$$[\theta_i \mid X_i, \eta = \hat{\eta}] \propto [X_i \mid \theta_i][\theta_i \mid \eta = \hat{\eta}]$$

These methods achieve a compromise between…
…"all the same" and "all completely unconnected".
...low variance/high bias and high variance/low bias.

Diagnostic test example: **What if you do not know the "prior" (prevalence)?**

Let a = sensitivity = 0.9, b = specificity = 0.8, n = #patients = 100.
Let T = $X_1 + \ldots + X_n$ = # of positive test results = 30**.**
The true disease prevalence is probably NOT 30/100!

(1) Standard Bayes requires that the prevalence is known. Suppose [disease]=0.10.
PPV = [disease | positive] = [positive|disease] x [disease] / [positive]
= 0.9x0.10 / (0.9x0.10+(1–0.8)x(1–0.10) = 0.09/(0.09+0.18) = **1/3**.

(2) Hierarchical Bayes requires a prior for the prevalence.
Suppose η = pr(disease)~Beta(1,9), and **we assume that the sensitivity and specificity do not vary as η varies**. Then the same formula applies, and [disease] = E([disease] | η] = E(η) = 1/(1+9) = 0.10 as before.

(3) Empirical Bayes directs us to estimate first the prevalence, then each disease status:

"Empirical": Use the marginal likelihood to estimate the prevalence:

$$[T \mid \eta] \propto \prod_{i=1}^{n} [X_i \mid \eta] = \boxed{\sum_{\theta_1,\ldots,\theta_n} \prod_{i=1}^{n} [X_i \mid \theta_i][\theta_i \mid \eta] = \prod_{i=1}^{n} \sum_{\theta_i=0}^{1} [X_i \mid \theta_i][\theta_i \mid \eta]}$$

$$= \prod_{i=1}^{n} \left(\sum_{\theta_i=0}^{1} [positive \mid \theta_i][\theta_i \mid \eta] \right)^{X_i} \left(\sum_{\theta_i=0}^{1} [negative \mid \theta_i][\theta_i \mid \eta] \right)^{1-X_i}$$

$$= \left(\sum_{\theta_i=0}^{1} [positive \mid \theta_i][\theta_i \mid \eta] \right)^{T} \left(\sum_{\theta_i=0}^{1} [negative \mid \theta_i][\theta_i \mid \eta] \right)^{n-T}$$

$$= \left(a\eta + (1-b)(1-\eta)\right)^{T} \left((1-a)\eta + b(1-\eta)\right)^{n-T}$$

Then the MLE and Method of Moments (both classical methods) agree:

$$a\hat{\eta} + (1-b)(1-\hat{\eta}) = T/n = 0.3$$

$$\hat{\eta} = \frac{T/n - (1-b)}{a-(1-b)} = \frac{0.3-0.2}{0.9-0.2} = 1/7 = 0.14$$

as our estimate of the prevalence.
Gut-check: Why is this estimate of prevalence lower than T/n? Could it be negative?

"Bayes": We get the predictive values (pos, neg) by a frequentist "plug in", $\hat{\eta}$ for η.

Positive predictive value = $\widehat{[disease \mid P]}$

$$= \frac{[P \mid disease]\widehat{[disease]}}{\widehat{[P]}} = \frac{a\hat{\eta}}{a\hat{\eta} + (1-b)(1-\hat{\eta})} = \frac{0.9x7^{-1}}{0.9x7^{-1} + 0.2x6x7^{-1}} = \frac{9}{9+12} = \frac{3}{7}.$$

Negative predictive value = $\widehat{[healthy \mid N]}$

$$= \frac{[N \mid healthy]\widehat{[healthy]}}{\widehat{[N]}} = \frac{b(1-\hat{\eta})}{b(1-\hat{\eta}) + (1-a)\hat{\eta}} = \frac{0.8x6x7^{-1}}{0.8x6x7^{-1} + 0.1x7^{-1}} = \frac{48}{48+1} = 0.98.$$

Key Empirical Bayes Concepts

Borrowing strength
Shrinking to the overall mean.
Reducing variance by increasing bias.
Plugging in.

What if we had not "borrowed strength"? This means, for example, estimating the disease status by maximizing the likelihood for fixed data:

For $X_i = Positive$, $[P \mid D] = 0.9$, $[P \mid \bar{D}] = 0.2$, so the "MLE" for θ_i is *D*.

But instead we used the data from the other *n*–1 observations.

Extensions

In real life, the sensitivity and specificity are not known perfectly,
the sensitivity and specificity may vary in different groups of patients,
the test result may be quantitative,
the quantitative test result may be informative about disease severity,
the quantitative test result may be informative about the loss function .

All these can be handled within the paradigm.

Example: Normal-normal sampling

Suppose $\theta_i \sim N(0,\tau)$ $i = 1,...,n,$ $X_i \mid \theta_i \sim N(\theta_i, \sigma_i)$ $i = 1,...,n,$ with independent sampling.

We suppose that τ is unknown, and (for starters) the σ_i are known.

(Each X_i could be a mean of n_i i.i.d. observations that share a common mean.)

Step 1: Empirical:

$X_i \mid \tau \sim N(0, \tau + \sigma_i)$ i.i.d., so $X_i^2 (\tau + \sigma_i)^{-1} \sim Chisq(1).$

Options for estimating τ include MLE and "method of moments" (but which moment?).

NOTES: (1) When "method of moments" matches $\mathrm{E}(1/\Sigma X_i^2)$ or $\mathrm{E}(1/\Sigma(X_i - \bar{X})^2)$, (inverse gamma) we get the famous James-Stein estimator, which showed that the normal mean ($n \geq 3$) is INADMISSIBLE under squared error loss.

(2) "Method of moments" matches the MARGINAL variation; subtracting to get $\hat{\tau}$ can lead to $\hat{\tau} < 0$. So we then set $\hat{\tau} = 0$.

(3) You don't have to assume that the prior mean is zero.

Step 2: Bayes: (or "Bayes-like")

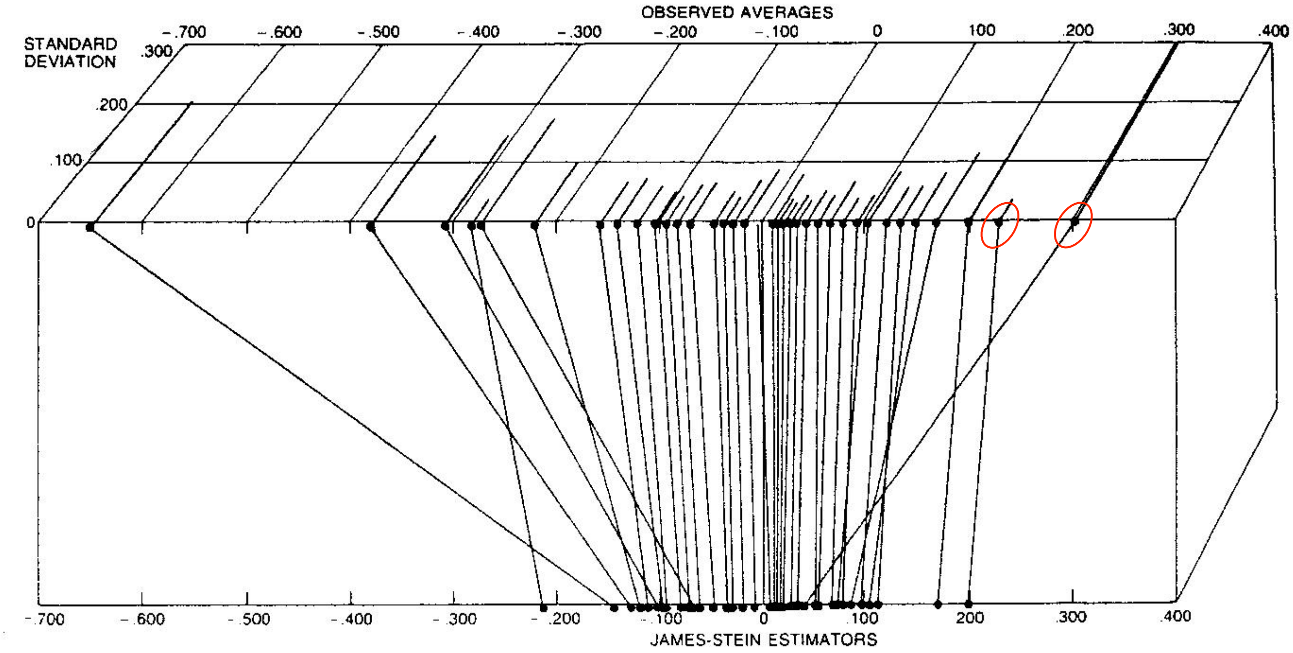
If we knew τ, then $\theta_i \mid X_i \sim N(\mu_i^*, \tau_i^*)$ $i = 1,...,n.$ where

$m_i^* = B_i X_i + (1 - B_i)0 = B_i X_i$, $\tau_i^* = B_i \sigma_i = (1 - B_i)\tau$ and $B_i = \tau / (\tau + \sigma_i)$ is the shrinker.

This is the linear regression of θ_i on X_i.

Instead, we plug in $\hat{\tau}$ for τ.

A famous example from Efron-Morris Scientific American 1977 ‘Stein's paradox in statistics’



SHRINKING of the observed toxoplasmosis rates to yield a set of James-Stein estimators substantially alters the apparent distribution of the disease. The shrinking factor is not the same for all the cities but instead depends on the standard deviation of the rate measured in that city. A large standard deviation implies that a measurement is based on a small sample and is subject to large random fluctuations; that measurement is therefore compressed more than the others are. In the El Salvador data the most extreme observations tend to be correlated with the largest standard deviations, again suggesting the unreliability of those measurements. Compared with the observed rates, the James-Stein estimators can be proved to have a smaller total error of estimation. They also provide a more accurate ranking of the cities.

The Empirical Bayes Confidence Interval Conundrum

No estimate is complete without a measure of accuracy.
For Empirical Bayes, how do you define a confidence interval for θ_I?

What should we mean by “coverage probability”? Should it be

- conditional on the data, averaged over the unknown θ’s?
- conditional on the unknown θ’s, averaged over the data?
- averaged over both the unknown θ’s and the data?

And, should it

- assume that $\eta = \hat{\eta}$?
- take into account our uncertainty about η?

Pointers to some answers:

Carl Morris
Nan Laird and Tom Louis.

Note that the ***Bayesian*** answer is straightforward: condition on the data, marginalize over things you don’t care about (here, η), look at the posterior distribution for each θ_I.
It turns out that Monte Carlo Markov Chain (MCMC) provides powerful machinery for this.