Bayesian Hierarchical Modeling

Tutorial 1: ISBA 2000, Crete

28 May 2000, 1–2.30pm

David Draper

Statistics Group

Department of Mathematical Sciences

University of Bath, England

`d.draper@maths.bath.ac.uk`

`http://www.bath.ac.uk/~masdd`

Based on the book of the same name, to be published by Springer-Verlag in 2000. **NB** Draft copies of Chapters 1 and 2 of this book are available in `PostScript` format for downloading from my web page, together with a variety of data sets and programs in `S+`, `C`, and `Maple`.

1. Introduction

This tutorial provides a very brief introduction to the formulation, fitting, and checking of $\boxed{\textbf{hierarchical}}$ or **multilevel** models from the Bayesian point of view.

Hierarchical models (**HMs**) arise frequently in **five main kinds of applications**:

$\boxed{\mathbf{1}}$ HMs are common in fields such as health and education, in which data—both outcomes and predictors—are often gathered in a **nested** or **hierarchical** fashion, e.g., patients within hospitals, or students within classrooms within schools.

HMs are thus also ideally suited to the wide range of applications in government and business in which single- or multi-stage **cluster samples** are routinely drawn, and offer a unified approach to the analysis of **random-effects** (variance-components) and **mixed models**.

Introduction (continued)

$\boxed{2}$ A different kind of nested data arises in **meta-analysis** in, e.g., medicine and the social sciences.

In this setting the goal is **combining information** from a number of studies of essentially the same phenomenon, to produce more accurate inferences and predictions than those available from any single study.

Here the data structure is subjects within studies, and as in the clustered case above there will generally be **predictors available at both the subject and study levels**.

$\boxed{3}$ When individuals—in medicine, for instance—are sampled cross-sectionally but then studied **longitudinally**, with outcomes observed at multiple time points for each person, a hierarchical data structure of the type studied in **repeated-measures** or **growth curve** analyses arises, with the readings at different time points nested within person.

Introduction (continued)

$\boxed{4}$ For simplicity people often try to model data as **(conditionally) IID** at a fairly high level of aggregation—for instance, by pretending that all the subjects in a sampling experiment are drawn **homogeneously** from a single population.

In fact, **heterogeneity** is often the rule rather than the exception, and frequently the available predictor variables do not "explain" this heterogeneity sufficiently.

With recent computational advances it is becoming increasingly straightforward to at least **describe** such heterogeneity with **mixture models** that employ **latent variables** (unobserved predictors) in a hierarchical structure.

Examples include **density estimation** with an unknown number of sub-populations mixed together and **Bayesian nonparametric modeling**, in which people work with distributions whose sample spaces are themselves sets of distributions instead of (say) real numbers.

Introduction (continued)

$\boxed{\mathbf{5}}$ Hierarchical modeling also provides a natural way to treat issues of **model selection** and **model uncertainty** with all types of data, not just cluster samples.

For example, in regression, if the data appear to exhibit residual variation that changes with the predictors, you can **expand the model** that assumes constant variation, by embedding it in a family of models that span a variety of assumptions about residual variation.

In this way, instead of having to choose one of these models and risk making the wrong choice, **you can work with several (or many) models at once**, weighting them in proportion to their plausibility given the data.

Introduction (continued)

In studying HMs two kinds of **technical** issues also arise:

- fully Bayesian computation in HMs requires the use of **simulation** methods such as those based on **Markov Chain Monte Carlo (MCMC)** ideas, and

- as usual with any class of statistical models there are questions of model (as opposed to MCMC) **diagnostics**.

MCMC methods will be covered in more detail in the second tutorial by Brad Carlin.

In this first session I will give a brief overview of some of the topics above, concentrating on two real examples: a **meta-analysis** of teacher expectancy studies in education, and a **random-effects Poisson regression** problem arising from a controlled trial of in-home geriatric assessment for elderly people in Scandinavia.

1.1 A Ten-Minute Review of de Finetti-style Bayesian Inference

$\boxed{\textbf{Case Study 1.}}$ Suppose you are interested in measuring the **quality of care** (e.g., Kahn et al., 1990) offered by one particular hospital.

(I am thinking of the **Royal United Hospital (RUH)** in Bath; you would no doubt have a **different** hospital in mind.)

As part of this you decide to examine the medical records of all patients treated at the RUH in one particular time window, say **January 2000–December 2003**, for one particular medical condition for which there is a strong *process-outcome link*, say **acute myocardial infarction (AMI; heart attack)**.

(*Process* is what health care providers do on behalf of patients; *outcomes* are what happens as a result of that care.)

In the **time window** you're interested in there will be about $\boxed{n = 400}$ AMI patients at the RUH.

Bayesian Modeling

To keep things simple let's ignore process for the moment and focus here on one particular outcome: **death status (mortality)** as of 30 days from hospital admission, coded 1 for dead and 0 for alive.

(In addition to process this will also depend on the **sickness at admission** of the AMI patients, but let's ignore that initially too.)

From the vantage point of December 1999, say, **what may be said** about the roughly 400 1s and 0s you will observe in 2000–03?

As a (de Finetti-style) Bayesian in this situation, your job is to quantify your uncertainty about these **observables**, i.e., your initial modeling task is **predictive** rather than inferential.

Unlike the frequentist approach to modeling, there is no samples-and-populations story here (these 400 people are **not** (literally) **a sample from any population**), so random variables do not arise in an attempt to describe the **process** of observing the repeatable event of random sampling.

Bayesian Modeling (continued)

Instead, random variables arise in a different way: quantifying your uncertainty (for the purpose of betting with someone about some aspect of the 1s and 0s, say) requires **eliciting** from yourself a joint probability distribution that **accurately** captures your judgments about the observables you have not yet seen:

$$\boxed{P_{B:\text{you}}(Y_1 = y_1, \ldots, Y_n = y_n)}.$$

Eliciting a 400-dimensional distribution doesn't sound easy; major **simplification** is evidently needed.

In this case, and many others, this is provided by **exchangeability** considerations.

If you have no relevant information that distinguishes one AMI patient from another, your uncertainty about the 400 1s and 0s is **symmetric**, in the sense that a random permutation of the **order** in which the 1s and 0s were labeled from 1 to 400 would leave your uncertainty about them unchanged.

Exchangeability

de Finetti (1930, 1964) called random variables with this property **exchangeable**:

> $\{Y_i, i = 1, \ldots, n\}$ are **exchangeable** if the distributions of $(Y_1, \ldots, Y_n)$ and $(Y_{\pi(1)}, \ldots, Y_{\pi(n)})$ are the same for all permutations $(\pi(1), \ldots, \pi(n))$.

NB Exchangeability and IID are not the same: IID implies exchangeability, and exchangeable Y_i do have identical marginal distributions, but they are not independent (if you were expecting *a priori* about 15% 1s, say (that's the 30-day death rate for AMI with average-quality care), the knowledge that in the first 50 outcomes 20 of them were deaths would certainly change your prediction of the 51st).

de Finetti also defined **partial** or **conditional** exchangeability (e.g., Draper et al., 1993): if, e.g., the gender X of the AMI patients were available, and there were evidence from the medical literature that 1s tended to be noticeably more likely for men than women, then you would probably want to assume *conditional* exchangeability of the Y_i given X (meaning that the male and female 1s and 0s, viewed as separate collections of random variables, are each unconditionally exchangeable). This is related to Fisher's (1956) idea of **recognizable subpopulations**.

de Finetti's Theorem for 1s and 0s

The judgment of exchangeability still seems to leave the joint distribution of the Y_i quite imprecisely specified. After defining the concept of exchangeability, however, de Finetti went on to prove a remarkable result: if you are willing to regard the $\{Y_i, i = 1, \ldots, n\}$ as part of an *infinite* exchangeable sequence of 1s and 0s (meaning that every finite subsequence is exchangeable), then there is a simple way to characterize your joint distribution, if it is to be internally consistent (e.g., de Finetti, 1975; Bernardo and Smith, 1994):

$\boxed{\textbf{de Finetti's Representation Theorem.}}$ If $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of 0–1 random quantities with probability measure P, there exists a distribution function $Q(\theta)$ such that the joint distribution $p(y_1, \ldots, y_n)$ for $Y_1, \ldots, Y_n$ is of the form

$$p(y_1, \ldots, y_n) = \int_0^1 \prod_{i=1}^n \theta^{y_i} (1-\theta)^{1-y_i} \, dQ(\theta) \ , \quad (1)$$

where $Q(\theta) = \lim_{n \to \infty} P(\frac{1}{n} \sum_{i=1}^n y_i \le \theta)$ and $\theta \stackrel{P}{=} \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n y_i$.

The distribution function Q will generally be well-behaved enough to have a density: $dQ(\theta) = p(\theta) d\theta$. In this case de Finetti's Theorem says

$$\begin{aligned} p(y_1, \ldots, y_n) &= \int_0^1 p(y_1, \ldots, y_n, \theta) d\theta \\ &= \int_0^1 p(y_1, \ldots, y_n | \theta) \, p(\theta) d\theta \qquad (2) \\ &= \int_0^1 \prod_{i=1}^n \theta^{y_i} (1-\theta)^{1-y_i} \, p(\theta) d\theta \ , \end{aligned}$$

The Simplest Hierarchical Model

In other words, it is **as if** (a) there were a **random** quantity called θ, interpretable as the **limiting relative frequency of 1s**, (b) conditional on this θ the Y_i are $\boxed{\text{IID B}(\theta)}$, and (c) θ itself has a distribution with density $p(\theta)$.

In yet other words, for a Bayesian whose uncertainty about the Y_i is **exchangeable**, the model for 0/1 outcomes may effectively be taken to be

$$\boxed{\begin{array}{rcl} \theta & \sim & p(\theta) \\ (Y_i|\theta) & \stackrel{\text{IID}}{\sim} & \text{B}(\theta), \ i = 1, \ldots, n. \end{array} \quad (3)}$$

<u>**NB**</u> (Thus exchangeability $\leftrightarrow$ *conditionally* IID given one or more parameters.) This is an example of the simplest kind of **hierarchical model (HM)**: a model at the top level for the underlying death rate θ, and then a model below that for the 0/1 mortality indicators Y_i conditional on θ.

Notice that to make sense of this in the Bayesian approach **you have to think of θ as a random variable**, even though logically it's a fixed unknown constant.

Bayes' Theorem

$p(\theta)$ does not involve $Y = (Y_1, \ldots, Y_n)$, so it must represent your uncertainty about θ before the data set Y arrives—it makes sense to call it your **prior distribution** for θ (although in the mathematics of de Finetti's Theorem it is just a **mixing distribution**).

Q: If $p(\theta)$ represents your uncertainty about θ before the data arrive, what represents this uncertainty *after* Y has been observed?

A: It has to be $p(\theta|Y)$, the conditional distribution for θ given how Y came out. It's natural to call this the **posterior distribution** for θ given Y.

Q: How do you get from $p(\theta)$ to $p(\theta|Y)$, i.e., how do you update your uncertainty about θ in light of the data?

A: Bayes' Theorem for continuous quantities:

$$p(\theta|Y) = \frac{p(\theta)\, p(Y|\theta)}{p(Y)}. \tag{4}$$

This requires some interpreting. As a Bayesian you are **conditioning on the data**, i.e., you are thinking of the left-hand side of (4) as a function of θ for fixed Y, so that must also be true of the right-hand side. Thus (a) $p(Y)$ is just a constant—in fact, you can think of it as the **normalizing constant**, put into the equation to make the product $p(\theta)\, p(Y|\theta)$ integrate to 1; and (b) $p(Y|\theta)$ may look like the usual frequentist sampling distribution for Y given θ (Bernoulli, in this case), but you have to think of it as a function of θ for fixed Y. When thought of this way—you could denote it $l(\theta|Y) = p(Y|\theta)$—Fisher (1922) called it the **likelihood function.**

Prior, Posterior, and Predictive Distributions

So Bayes' theorem becomes

$$
\begin{aligned}
p(\theta|Y) &= \qquad c \qquad \cdot\, p(\theta) \cdot l(\theta|Y) \ , \\
\text{posterior} &= \begin{pmatrix} \text{normalizing} \\ \text{constant} \end{pmatrix} \cdot \text{prior} \cdot \text{likelihood} \ .
\end{aligned}
$$

You can also readily construct **predictive distributions** for the Y_i before they are observed, or for future Y_i once some of them are known, e.g., the posterior predictive distribution for $(Y_{m+1}, \ldots, Y_n)$ given $(Y_1, \ldots, Y_m)$, i.e., $p(y_{m+1}, \ldots, y_n | y_1, \ldots, y_m)$ is just

$$
\begin{aligned}
&\int_0^1 p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)\, p(\theta | y_1, \ldots, y_m)\, d\theta \\
&\qquad = \int_0^1 p(y_{m+1}, \ldots, y_n | \theta)\, p(\theta | y_1, \ldots, y_m)\, d\theta \qquad (5) \\
&\qquad = \int_0^1 \prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}\, p(\theta | y_1, \ldots, y_m)\, d\theta \ .
\end{aligned}
$$

Notice an important simplification here in going from $p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)$ to $p(y_{m+1}, \ldots, y_n | \theta)$: conditional on θ the Y_i are independent, i.e., if you know θ the individual values of $Y_1, \ldots, Y_m$ will not help you to predict $Y_{m+1}, \ldots, Y_n$.

This also brings up a key difference between a parameter like θ on the one hand and the Y_i, before you have observed any data, on the other: parameters are inherently **unobservable**. This makes it harder to evaluate the quality of your uncertainty assessments about θ than to do so about the observable Y_i.

Computation: Conjugate Analysis

Prior specification in the mortality data. Let's say (a) you know that the 30-day AMI mortality rate given average care and average sickness at admission in England is about 15%, (b) You know little about care or patient sickness at the RUH, but (c) You would be somewhat surprised (on Central Limit Theorem grounds) if the "underlying rate" at the RUH was much less than 5% or more than 30% (note the asymmetry).

To quantify these judgments you seek a **flexible family of densities** on (0,1), one of whose members has mean 0.15 and (say) 95% central interval (0.05,0.30).

A convenient family for this purpose is the **beta** distributions, $\text{Be}(\theta|\alpha,\beta) = c\,\theta^{\alpha-1}(1-\theta)^{\beta-1}$, for two reasons:

- This family exhibits a **wide variety of distributional shapes** (e.g., Johnson and Kotz, 1970); and
- The likelihood in this problem comes from the Bernoulli sampling distribution for the Y_i, $p(y_1,\ldots,y_n|\theta) = l(\theta|Y) = \theta^S(1-\theta)^{n-S}$, where S is the sum of the Y_i. Thus with this choice of prior, the likelihood and prior (and thus the posterior) have the same distributional form, $\theta^r(1-\theta)^s$, which makes life computationally much easier. The beta family is thus **conjugate** to the Bernoulli/binomial likelihood.

Conjugate Analysis (continued)

Conjugate analysis—finding conjugate priors for standard likelihoods and restricting attention to them on tractability grounds—is one of only two fairly general methods (the other is **asymptotic analysis**; see Bernardo and Smith, 1994) for getting closed-form answers in the Bayesian approach.

In the mortality example, trial and error shows $\alpha = 4.5$ and $\beta = 25.5$ produce approximately the desired mean and central interval (this distribution has mode .125 and SD .064). α and β are called $\boxed{\textbf{hyperparameters}}$ since they are parameters of the prior distribution. Written **hierarchically** the model is

$$\begin{array}{rcll} (\alpha, \beta) & = & (4.5, 25.5) & \text{(hyperparameters)} \\ (\theta \mid \alpha, \beta) & \sim & \text{Be}(\alpha, \beta) & \text{(prior)} \\ (Y_1, \ldots, Y_n \mid \theta) & \stackrel{\text{IID}}{\sim} & \text{B}(\theta) & \text{(likelihood)} \end{array} \qquad (6)$$

The conjugacy of the prior leads to a simple closed form for the posterior here: with y as the vector of observed $Y_i, i = 1, \ldots, n$ and S as the sum of the Y_i (the sufficient statistic for θ with the Bernoulli likelihood),

$$\begin{aligned} p(\theta \mid y, \alpha, \beta) & = c \; p(y \mid \theta) \, p(\theta \mid \alpha, \beta) \\ & = c \; \theta^S (1-\theta)^{n-S} \, \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ & = c \; \theta^{(S+\alpha)-1} (1-\theta)^{(n-S+\beta)-1}, \end{aligned} \qquad (7)$$

i.e., the posterior for θ is $\text{Be}(\alpha + S, \beta + n - S)$. This gives the hyperparameters a nice interpretation in terms of **effective information content of the prior**: it is as if the data $(\text{Be}(S+1, n-S+1))$ were worth $(S+1) + (n-S+1) \doteq n$ observations and the prior $(\text{Be}(\alpha, \beta))$ were worth $(\alpha + \beta)$ observations. This can be used to judge whether the prior is "too informative"—here it is equivalent to $(4.5 + 25.5) =$ 30 binary observables with a mean of .15.

Conjugate Analysis (continued)

Suppose the $n = 400$ observed mortality indicators consist of $S = 72$ 1's and $(n - S) = 328$ 0's. Then the prior is Be(4.5,25.5), the likelihood is Be(73,329), the posterior for θ is Be(76.5,353.5), and the three densities plotted on the same graph come out as in Figure 1.1:

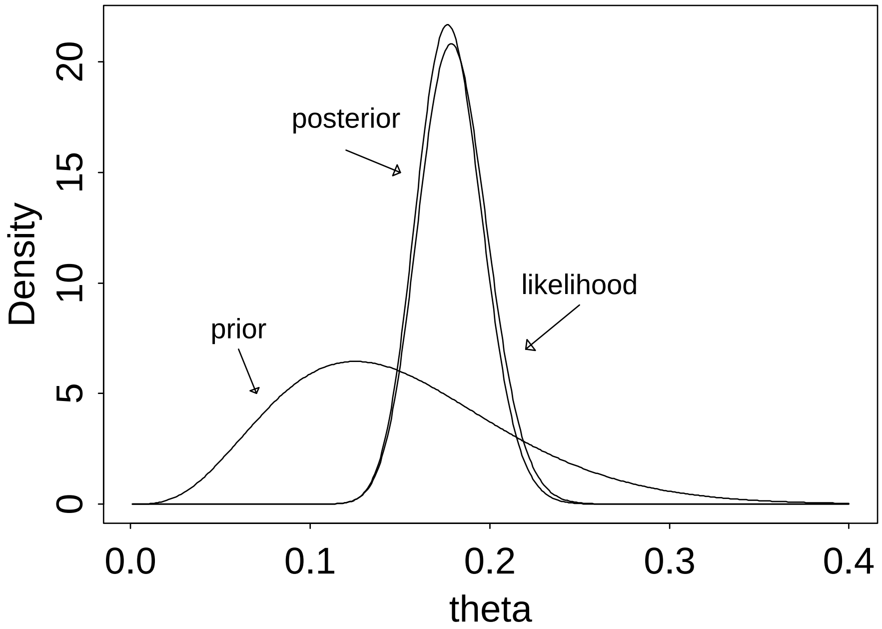

Figure 1.1. Prior, likelihood, and posterior distributions for the RUH mortality data.

In this case the posterior and the likelihood nearly coincide, because the data information outweighs the prior information by 400/30 = more than 13 to 1. The mean of a Be(α, β) distribution is $\alpha/(\alpha + \beta)$; with this in mind the posterior mean has a nice expression as a weighted average of the prior mean and data mean, with weights determined by the **effective sample size** of the prior, $(\alpha + \beta)$, and the data sample size n:

$$\frac{\alpha + S}{\alpha + \beta + n} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\left(\frac{\alpha}{\alpha + \beta}\right) + \left(\frac{n}{\alpha + \beta + n}\right)\left(\frac{S}{n}\right)$$

$$\begin{array}{c} \text{posterior} \\ \text{mean} \end{array} = \left(\begin{array}{c} \text{prior} \\ \text{weight} \end{array}\right)\left(\begin{array}{c} \text{prior} \\ \text{mean} \end{array}\right) + \left(\begin{array}{c} \text{data} \\ \text{weight} \end{array}\right)\left(\begin{array}{c} \text{data} \\ \text{mean} \end{array}\right)$$

$$.178 = (.070) \quad (.15) + (.93) \quad (.18) .$$

1.2 Continuous Outcomes

For **continuous outcomes** there is an analogue of de Finetti's Theorem that is equally central to Bayesian model-building (e.g., Bernardo and Smith, 1994):

de Finetti's Theorem for Continuous Outcomes. If $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of real-valued random quantities with probability measure P, there exists a probability measure Q over $\mathcal{D}$, the space of all distribution functions on R, such that the joint distribution function of $Y_1, \ldots, Y_n$ has the form

$$P(y_1, \ldots, y_n) = \int_{\mathcal{D}} \prod_{i=1}^{n} F(y_i)\, dQ(F), \qquad (8)$$

where $Q(F) = \lim_{n \to \infty} P(F_n)$ and F_n is the empirical distribution function based on $Y_1, \ldots, Y_n$.

In other words, **exchangeability of real-valued observables** is equivalent to the HM

$$\boxed{\begin{array}{rcll} F & \sim & p(F) & \text{(prior)} \\ (Y_1, \ldots, Y_n | F) & \stackrel{\text{IID}}{\sim} & F & \text{(likelihood)} \quad (9) \end{array}}$$

for some prior distribution p on the **set $\mathcal{D}$ of all possible distribution functions**.

This prior makes the continuous form of de Finetti's Theorem considerably harder to apply: to take the elicitation task seriously is to try to specify a measure on function space (F is in effect an **infinite-dimensional** parameter).

Continuous Outcomes (continued)

(**NB** This task is not unique to Bayesians—the analogous question in frequentist analyses of observational data is **"Where does the likelihood come from?"**)

What people often do in practice is to appeal to considerations that narrow down the field, such as an *a priori* judgment that the Y_i ought to be **symmetrically** distributed about a location parameter μ, and then try to use a fairly rich parametric family satisfying (e.g.) the symmetry restriction as a substitute for all of $\mathcal{D}$.

Strictly speaking you are not supposed to look at the Y_i while specifying your prior on $\mathcal{D}$—this can lead to a failure to fully assess and propagate **model uncertainty**—but not doing so can permit the data to surprise you in ways that would make you want to go back and revise your prior.

Two ways out of this dilemma are **out-of-sample predictive validation** and **Bayesian nonparametrics**, both of which I will briefly examine later.

2. Hierarchical Models for Combining Information

2.1 Formulating hierarchical models for quantitative outcomes from scientific context

Case Study 2: *Meta-analysis of effects of aspirin on heart attacks*. Table 2.1 (Draper et al., 1993a) gives the number of patients and mortality rate from all causes, for six randomized controlled experiments comparing the use of aspirin and placebo by patients following a heart attack.

Table 2.1. Aspirin meta-analysis data.

	Aspirin		Placebo	
Study (i)	# of Patients	Mortality Rate (%)	# of Patients	Mortality Rate (%)
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

	Comparison			
Study (i)	y_i = Diff (%)	$\sqrt{V_i}$ = SE of Diff (%)	$Z_i^{\ddagger}$	$p_i^{\S}$
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

$^{\ddagger}Z_i$ denotes the ratio of the difference in mortality rates over its standard error, assuming a binomial distribution. $^{\S}p_i$ is the one-sided p value associated with Z_i, using the normal approximation.

Meta-Analysis

The first five trials are reasonably consistent in showing a (weighted average) mortality decline for aspirin patients of 2.3 percentage points, a 20% drop from the (weighted average) placebo mortality of 11.5% (this difference is **highly clinically significant**). However, the sixth and largest trial, AMIS, went the other way: an *increase* of 1.2 percentage points in aspirin mortality (a 12% rise from the placebo baseline of 9.7%).

Some relevant questions (Draper, 1995): $\boxed{\mathbf{Q}_1}$ Why did AMIS get such different results? $\boxed{\mathbf{Q}_2}$ What should be done next to reduce the uncertainty about Q_1? $\boxed{\mathbf{Q}_3}$ If you were a doctor treating a patient like those eligible for the trials in Table 2.1, what therapy should you employ while answers to Q_1 and Q_2 are sought? One possible paraphrase of Q_3: $\boxed{\mathbf{Q}_4}$ How should the information from these six experiments be **combined** to produce a more informative summary than those obtained from each experiment by itself?

The discipline of $\boxed{\textbf{meta-analysis}}$ is devoted to answering questions like Q_4. One leading school of frequentist meta-analysis (e.g., Hedges and Olkin, 1985) looks for methods for combining the Z and p values in Table 2.1, an approach that often leads only to an overall p value. A more satisfying form of meta-analysis (which has both frequentist and Bayesian versions) builds a **hierarchical model (HM)** that indicates how to combine information from the mortality differences in the Table. A Gaussian meta-analysis model for the aspirin data, for example (Draper et al., 1993a), might look like

$$\begin{array}{cclc} (\theta, \sigma^2) & \sim & p(\theta, \sigma^2) & \text{(prior)} \\ (\theta_i | \theta, \sigma^2) & \stackrel{\text{IID}}{\sim} & N\big(\theta, \sigma^2\big) & \text{(underlying effects)} \\ (y_i | \theta_i) & \stackrel{\text{indep}}{\sim} & N(\theta_i, V_i) & \text{(data)} . \end{array} \qquad (10)$$

A Gaussian HM

The bottom level of (10), the **data** level of the HM, says that—because of relevant differences in patient cohorts and treatment protocols—each study has its own underlying treatment effect θ_i, and the observed mortality differences y_i are like random draws from a normal distribution with mean θ_i and variance V_i (the normality is reasonable because of the CLT, given the large numbers of patients). In meta-analyses of data like those in Table 2.1 the V_i are typically taken to be known (again because the patient sample sizes are so big), $V_i = SE_i^2$, where SE_i is the standard error of the mortality difference for study i in Table 2.1.

The middle level of the HM is where you would bring in the $\boxed{\textbf{study-level covariates}}$, if you have any, to try to explain why the studies differ in their underlying effects. Here there are no study-level covariates, so the middle level of (10) is equivalent to a Gaussian regression with no predictor variables. Why the "error" distribution should be Gaussian at this level of the HM is not clear—it's a conventional choice, not a choice that is automatically scientifically reasonable (in fact I will challenge it later). σ^2 in this model represents **study-level heterogeneity**.

The top level of (10) is where the **prior** distribution on the regression parameters from the middle level is specified. Here, with only an intercept term in the regression model, a popular conventional choice is a normal/scaled-inverse-χ^2 prior (see Gelman et al. 1995).

$\boxed{\textbf{Fixed effects and random effects.}}$ If σ^2 were known somehow to be 0, all of the θ_i would have to be equal deterministically to a common value θ, yielding a simpler model: $(y_i|\theta) \stackrel{\text{indep}}{\sim} N(\theta, V_i), \theta \sim p(\theta)$. Meta-analysts call this a **fixed-effects** model, and refer to model (1) as a **random-effects** model. When σ^2 is not assumed to be 0, with this terminology the θ_i are called **random effects**.

2.2 Approximate Fitting of Gaussian Hierarchical Models: Maximum Likelihood and Empirical Bayes

Fitting HM (10). Some algebra based on model (10) yields that the conditional distributions of the study-level effects θ_i given the data and the parameters (θ, σ^2), have a simple and revealing form:

$$(\theta_i | y_i, \theta, \sigma^2) \stackrel{\text{indep}}{\sim} N(\theta_i^*, V_i(1 - B_i)) \tag{11}$$

$$\text{with} \quad \theta_i^* = (1 - B_i)\, y_i + B_i\, \theta \quad \text{and} \quad B_i = \frac{V_i}{V_i + \sigma^2}. \tag{12}$$

In other words, the conditional mean of the effect for study i given y_i, θ, and σ^2 is a **weighted average** of the sample mean for that study, y_i, and the overall mean θ.

The weights are given by the so-called **shrinkage factors** B_i (e.g., Draper et al., 1993a), which in turn depend on how the variability V_i within study i compares to the between-study variability σ^2: the more accurately y_i estimates θ_i, the more weight the "local" estimate y_i gets in the weighted average. The term *shrinkage* refers to the fact that, with this approach, unusually high or low individual studies are drawn back or "shrunken" toward the overall mean in the calculation $(1 - B_i)\, y_i + B_i\, \theta$. Note that θ_i^* uses data from all the studies to estimate the effect for study i—this is referred to as **borrowing strength** in the estimation process.

Closed-form expressions for $p(\theta | y)$ and $p(\theta_i | y)$ with $y = (y_1, \ldots, y_k), k = 6$ are not available even with a normal/scaled-inverse-χ^2 prior for (θ, σ^2); a full Bayesian analysis of these data is most readily accomplished with Markov Chain Monte Carlo methods.

Maximum Likelihood and Empirical Bayes

In the meantime **maximum likelihood** calculations provide some idea of what to expect: the likelihood function based on model (1) is

$$l(\theta, \sigma^2 | y) = c \prod_{i=1}^{k} \frac{1}{\sqrt{V_i + \sigma^2}} \exp\left[-\frac{1}{2} \sum_{i=1}^{k} \frac{(y_i - \theta)^2}{V_i + \sigma^2} \right]. \qquad (13)$$

The maximum likelihood estimates (MLEs) $(\hat{\theta}, \hat{\sigma}^2)$ then turn out to be the **iterative** solutions to the following equations:

$$\hat{\theta} = \frac{\sum_{i=1}^{k} \hat{W}_i \, y_i}{\sum_{i=1}^{k} \hat{W}_i} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \hat{W}_i^2 \left[(y_i - \hat{\theta})^2 - V_i \right]}{\sum_{i=1}^{k} \hat{W}_i^2}, \qquad (14)$$

$$\text{where} \quad \hat{W}_i = \frac{1}{V_i + \hat{\sigma}^2}. \qquad (15)$$

Start with $\hat{\sigma}^2 = 0$ and iterate (5–6) to convergence (if $\hat{\sigma}^2$ converges to a negative value, $\hat{\sigma}^2 = 0$ is the MLE). The MLEs of the θ_i are then given by

$$\hat{\theta}_i = (1 - \hat{B}_i) \, y_i + \hat{B}_i \, \theta \quad \text{where} \quad \hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2}. \qquad (16)$$

These are called **empirical Bayes** (EB) estimates of the study-level effects, because it turns out that this analysis approximates a fully Bayesian solution by (in effect) using the data to estimate the prior specifications for θ and σ^2.

Large-sample (mainly meaning large k) approximations to the (frequentist) distributions of the MLEs are given by

$$\hat{\theta} \sim N\left(\theta, \left[\sum_{i=1}^{k} \frac{1}{V_i + \hat{\sigma}^2} \right]^{-1} \right) \quad \text{and} \quad \hat{\theta}_i \sim N\big(\theta_i, V_i(1 - \hat{B}_i)\big). \qquad (17)$$

MLEB (continued)

NB The variances in (17) *do not account fully for the uncertainty in* σ^2 and therefore underestimate the actual sampling variances for small k (adjustments are available; see, e.g., Morris (1983, 1988)).

MLEB estimation can be implemented simply in about 15 lines of `S-PLUS` code (Table 2.2).

Table 2.2. `S-PLUS` program to perform MLEB calculations.

```
mleb <- function( y, V, m ) {
  sigma.2 <- 0.0
  for ( i in 1:m ) {
    W <- 1.0 / ( V + sigma.2 )
    theta <- sum( W * y ) / sum( W )
    sigma.2 <- sum( W^2 * ( ( y - theta )^2 - V ) ) / sum( W^2 )
    B <- V / ( V + sigma.2 )
    effects <- ( 1 - B ) * y + B * theta
    se.theta <- 1.0 / sqrt( sum( 1.0 / ( V + sigma.2 ) ) )
    se.effects <- sqrt( V * ( 1.0 - B ) )
    print( c( i, theta, se.theta, sigma.2 ) )
    print( cbind( W, ( W / sum( W ) ), B, y, effects, se.effects ) )
  }
}
```

With the aspirin data it takes 18 iterations (less than 0.1 second on a 400MHz `UltraSPARC` Unix machine) to get 4-digit convergence to the summaries in Table 2.3 and the following estimates (standard errors in parentheses):

$$\hat{\theta} = 1.45 \ (0.809), \quad \hat{\sigma}^2 = 1.53.$$

Table 2.3. Maximum likelihood empirical Bayes meta-analysis of the aspirin data.

study(i)	$\hat{W}_i$	normalized $\hat{W}_i$	$\hat{B}_i$	y_i	$\hat{\theta}_i$	$\widehat{SE}(\hat{\theta}_i)$
1	0.235	0.154	0.640	2.77	1.92	0.990
2	0.308	0.202	0.529	2.50	1.94	0.899
3	0.143	0.0934	0.782	1.84	1.53	1.09
4	0.232	0.151	0.646	2.56	1.84	0.994
5	0.183	0.120	0.719	2.31	1.69	1.05
6	0.427	0.280	0.346	−1.15	−0.252	0.728

Aspirin Meta-Analysis: Conclusions

Note that (1) AMIS gets much less weight (normalized $\hat{W}_i$) than would have been expected given its small V_i; (2) the shrinkage factors ($\hat{B}_i$) are considerable, with AMIS shrunk almost all the way into positive territory (see Figure 2.1); (3) there is **considerable study-level heterogeneity** ($\hat{\sigma} \doteq 1.24$ percentage points of mortality); and (4) the standard errors of the effects are by and large smaller than the $\sqrt{V_i}$ (from the borrowing of strength) but are still considerable.

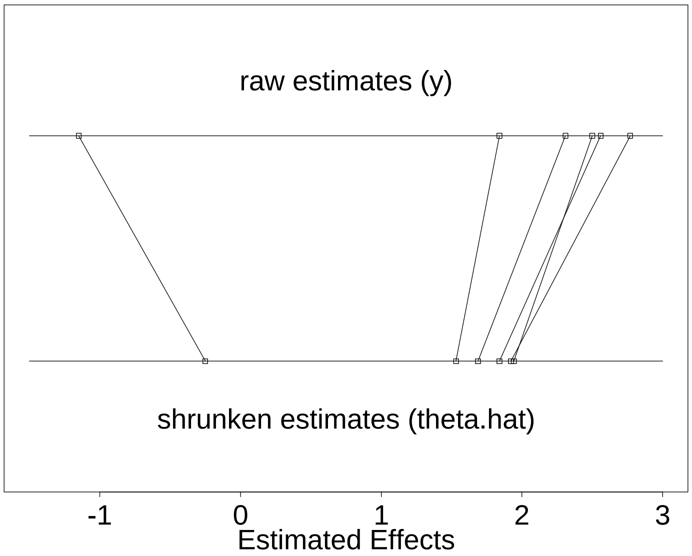

Figure 2.1. **Shrinkage plot** for the aspirin MLEB meta-analysis.

The 95% interval estimate of θ, the overall underlying effect of aspirin on mortality, from this approach comes out

$$\hat{\theta} \pm 1.96 \cdot \widehat{SE}(\hat{\theta}) \doteq (-0.140, 3.03),$$

which if interpreted Bayesianly gives

$$\boxed{P(\text{aspirin reduces mortality}|\text{data}) \doteq 1 - \Phi\left(\frac{0-1.45}{0.809}\right) = \mathbf{0.96}},$$

where Φ is the standard normal CDF. Thus although the interval includes 0, so that in a frequentist sense the effect is not statistically significant, **in fact from a Bayesian point of view the evidence is running strongly in favor of aspirin's usefulness**.

Educational Meta-Analysis

In many cases (as with this example) empirical Bayes methods have the advantage of yielding closed-form solutions, but I view them at best as approximations to fully Bayesian analyses—which can in any case be carried out with MCMC methods—so I won't have any more to say about EB methods here (see Carlin and Louis, 1996, for more on this topic).

2.3 Incorporating Study-Level Covariates

Case Study: *Meta-analysis of the effect of teacher expectancy on student IQ* (Bryk and Raudenbush, 1992). Do teachers' expectations influence students' intellectual development, as measured by IQ scores?

Table 2.4. Results from 19 experiments estimating the effects of teacher expectancy on pupil IQ.

Study (i)	Weeks of Prior Contact (x_i)	Estimated Effect Size (y_i)	Standard Error of $y_i = \sqrt{V_i}$
1. Rosenthal et al. (1974)	2	0.03	0.125
2. Conn et al. (1968)	3	0.12	0.147
3. Jose & Cody (1971)	3	-0.14	0.167
4. Pellegrini & Hicks (1972)	0	1.18	0.373
5. Pellegrini & Hicks (1972)	0	0.26	0.369
6. Evans & Rosenthal (1969)	3	-0.06	0.103
7. Fielder et al. (1971)	3	-0.02	0.103
8. Claiborn (1969)	3	-0.32	0.220
9. Kester & Letchworth (1972)	0	0.27	0.164
10. Maxwell (1970)	1	0.80	0.251
11. Carter (1970)	0	0.54	0.302
12. Flowers (1966)	0	0.18	0.223
13. Keshock (1970)	1	-0.02	0.289
14. Henrickson (1970)	2	0.23	0.290
15. Fine (1972)	3	-0.18	0.159
16. Greiger (1970)	3	-0.06	0.167
17. Rosenthal & Jacobson (1968)	1	0.30	0.139
18. Fleming & Anttonen (1971)	2	0.07	0.094
19. Ginsburg (1970)	3	-0.07	0.174

Teacher Expectancy

Raudenbush (1984) found $k = 19$ experiments, published between 1966 and 1974, estimating **the effect of teacher expectancy on student IQ** (Table 2.4). In each case the experimental group was made up of children for whom teachers were (deceptively) encouraged to have high expectations (e.g., experimenters gave treatment teachers lists of students, actually chosen at random, who allegedly displayed dramatic potential for intellectual growth), and the controls were students about whom no particular expectations were encouraged.

The estimated $\boxed{\textbf{effect sizes}}$ $y_i = (\bar{T}_i - \bar{C}_i)/\text{SD}_{i:\text{pooled}}$ (column 3 in Table 2.4) ranged from $-.32$ to $+1.18$; why? One good reason: the studies differed in how well the experimental teachers knew their students at the time they were given the deceptive information—this time period x_i (column 2 in Table 2.4) ranged from 0 to 3 weeks. Figure 2.2 plots y_i against x_i—you can see that **the studies with bigger x_i had smaller IQ effects on average**.

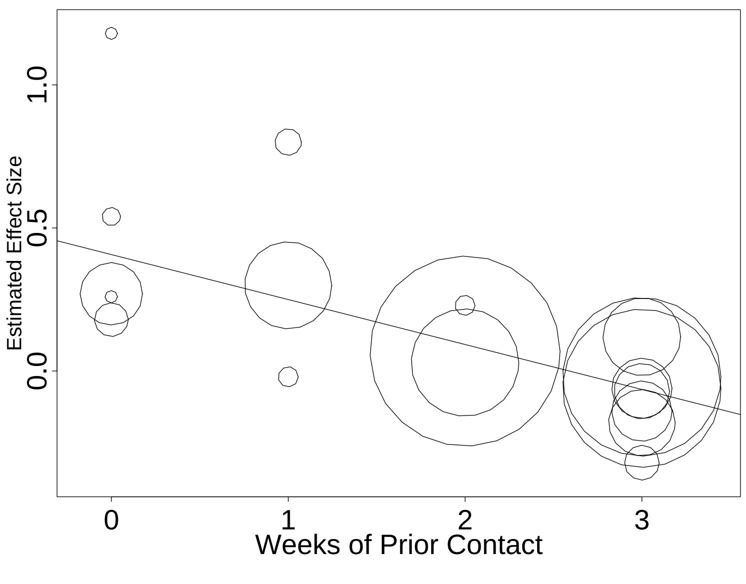

Figure 2.2. Scatterplot of estimated effect size against weeks of prior contact in the IQ meta-analysis. Radii of circles are proportional to $w_i = V_i^{-1}$ (see column 4 in Table 2.4); fitted line is from weighted regression of y_i on x_i with weights w_i.

Conditional Exchangeability

Evidently model (1) will not do here—it says that your predictive uncertainty about all the studies is *exchangeable* (similar, i.e., according to (1) the underlying study-level effects θ_i are like IID draws from a normal distribution), whereas Figure 2.2 clearly shows that the x_i are useful in predicting the y_i. This is another way to say that your uncertainty about the studies is not unconditionally exchangeable but $\boxed{\textbf{conditionally exchangeable given } x}$ (Draper et al., 1993b).

In fact Figure 2.2 suggests that the y_i (and therefore the θ_i) are related **linearly** to the x_i. Bryk and Raudenbush, working in the frequentist paradigm, fit the following HM to these data:

$$
\begin{array}{rcll}
(\theta_i | \, \alpha, \beta, \sigma^2) & \stackrel{\text{indep}}{\sim} & N\big(\alpha + \beta \, x_i, \sigma^2\big) & \text{(underlying effects)} \\
(y_i | \, \theta_i) & \stackrel{\text{indep}}{\sim} & N(\theta_i, V_i) & \text{(data).} \quad (18)
\end{array}
$$

According to this model the estimated effect sizes y_i are like draws from a Gaussian with mean θ_i and variance V_i, the squared standard errors from column 4 of Table 2.4—here as in model (10) the V_i are taken to be known—and the θ_i themselves are like draws from a Gaussian with mean $\alpha + \beta x_i$ and variance σ^2.

The top level of this HM in effect assumes, e.g., that the 5 studies with $x = 0$ are sampled representatively from {all possible studies with $x = 0$}, and similarly for the other values of x. This (and the Gaussian choice on the top level) are **conventional assumptions, not automatically scientifically reasonable**—for example, if you know of some way in which (say) two of the studies with $x = 3$ differ from each other that is relevant to the outcome of interest, then you should include this in the model as a study-level covariate along with x.

An MLEB Drawback

Bryk and Raudenbush used MLEB methods, based on the **EM algorithm**, to fit this model. As in Section 2.2, this estimation method combines the two levels of model (18) to construct a single likelihood for the y_i, and then maximizes this likelihood as usual in the ML approach. They obtained $(\hat{\alpha}, \hat{\beta}) = (.407 \pm .087, -.157 \pm .036)$ and $\hat{\sigma}^2 = 0$, naively indicating that all of the study-level variability has been "explained" by the covariate x.

However, from a Bayesian point of view, this model is missing a third layer:

$$\begin{aligned} (\alpha, \beta, \sigma^2) &\sim p(\alpha, \beta, \sigma^2) \\ (\theta_i | \alpha, \beta, \sigma^2) &\stackrel{\text{IID}}{\sim} N\big(\alpha + \beta(x_i - \bar{x}), \sigma^2\big) \\ (y_i | \theta_i) &\stackrel{\text{indep}}{\sim} N(\theta_i, V_i). \end{aligned} \qquad (19)$$

(it will help convergence of MCMC methods to make α and β uncorrelated by $\boxed{\textbf{centering}}$ the x_i at 0 rather than at $\bar{x}$). As will subsequently become clear, the trouble with MLEB is that in Bayesian language **it assumes in effect that the posterior for σ^2 is point-mass on the MLE**. This is bad (e.g., Morris, 1983) for two reasons:

- If the posterior for σ^2 is highly skewed, the mode will be a poor summary; and

- Whatever point-summary you use, pretending the posterior SD for σ^2 is zero fails to propagate uncertainty about σ^2 through to uncertainty about α, β, and the θ_i.

The best way to carry out a fully Bayesian analysis of model (19) is with **Markov Chain Monte Carlo** methods.

3. Hierarchical Model Selection

Case Study: *In-home geriatric assessment (IHGA)*. In an experiment conducted in the 1980s (Hendriksen et al. 1986), 572 elderly people living in a number of villages in Denmark were randomized, 287 to a **control** (C) group (who received standard care) and 285 to an **experimental** (E) group (who received standard care plus IHGA: a kind of preventive medicine in which each person's medical and social needs were assessed and acted upon individually).

One important outcome was the number of **hospitalizations** during the two-year life of the study (Table 3.1).

Table 3.1. Distribution of number of hospitalizations in the IHGA study over a two-year period.

Group	Number of Hospitalizations 0	1	2	3	4	5	6	7	n	Mean	SD
Control	137	77	46	12	8	4	0	2	287	0.944	1.24
Experimental	147	83	37	13	3	1	1	0	285	0.768	1.01

Evidently IHGA lowered the mean hospitalization rate (for these elderly Danish people, at least) by $(0.944 - 0.768) = \mathbf{0.176}$, which is about a $100\left(\frac{0.944-0.768}{0.768}\right) = \mathbf{19\%}$ reduction from the control level, a difference that is **large in clinical terms**.

Modeling the IHGA Data

An off-the-shelf analysis of this experiment might pretend (**Model 0**) that the data are Gaussian,

$$\begin{aligned} \left(C_i|\mu_C,\sigma_C^2\right) &\sim N\left(\mu_C,\sigma_C^2\right), i=1,\ldots,n_C, \\ \left(E_j|\mu_E,\sigma_E^2\right) &\sim N\left(\mu_E,\sigma_E^2\right), j=1,\ldots,n_E, \end{aligned} \quad (20)$$

and use the ordinary two-independent-samples "z-machinery":

```
loki $ S

S-PLUS : Copyright (c) 1988, 1998 MathSoft, Inc.
S : Copyright Lucent Technologies, Inc.
Version 5.0 Release 3 for Sun SPARC, SunOS 5.5 : 1998
Working data will be in .Data

> C <- c( rep( 0, 138 ), rep( 1, 77 ), rep( 2, 46 ),
    rep( 3, 12 ), rep( 4, 8 ), rep( 5, 4 ), rep( 7, 2 ) )

> print( n.C <- length( C ) )

[1] 287                  # sample size in the control group

> mean( C )

[1] 0.9442509            # control group mean

> sqrt( var( C ) )

[1] 1.239089             # control group standard deviation
                         #   (SD)
> table( C )

  0  1  2  3 4 5 7       # control group frequency
138 77 46 12 8 4 2       #   distribution
```

Analysis of Model 0

```
> E <- c( rep( 0, 147 ), rep( 1, 83 ), rep( 2, 37 ),
    rep( 3, 13 ),rep( 4, 3 ), rep( 5, 1 ), rep( 6, 1 ) )

> print( n.E <- length( E ) )

[1] 285                 # sample size in the experimental
                        #   group
> mean( E )

[1] 0.7684211             # experimental group mean

> sqrt( var( E ) )

[1] 1.008268              # experimental group SD

> table( E )

   0  1  2  3 4 5 6       # experimental group
 147 83 37 13 3 1 1       # frequency distribution

> print( effect <- mean( E ) - mean( C ) )

[1] -0.1758298            # mean difference ( E - C )

> effect / mean( C )

[1] -0.1862109            # relative difference ( E - C ) / C

> print( SE.effect <- sqrt( var( C ) / n.C + var( E ) / n.E ) )

[1] 0.09442807            # standard error of the difference

> print( CI <- c( effect - 1.96 * SE.effect,
    effect + 1.96 * SE.effect ) )

[1] -0.3609089 0.009249204 # the 95% confidence interval from
                           # model 0 runs from -.36 to +.01
```

Deficiencies of Model 0

The frequentist analysis of Model 0 is equivalent to a Bayesian analysis of the same model with **diffuse** priors on the control and experimental group means and SDs $(\mu_C, \sigma_C, \mu_E, \sigma_E)$, and is summarized in Table 3.2.

Table 3.2. Summary of analysis of Model 0.

	Posterior		
	Mean	SD	95% Interval
Treatment effect $(\mu_E - \mu_C)$	-0.176	0.0944	$(-0.361, 0.009)$

However, both distributions have long right-hand tails; in fact they look rather **Poisson**.

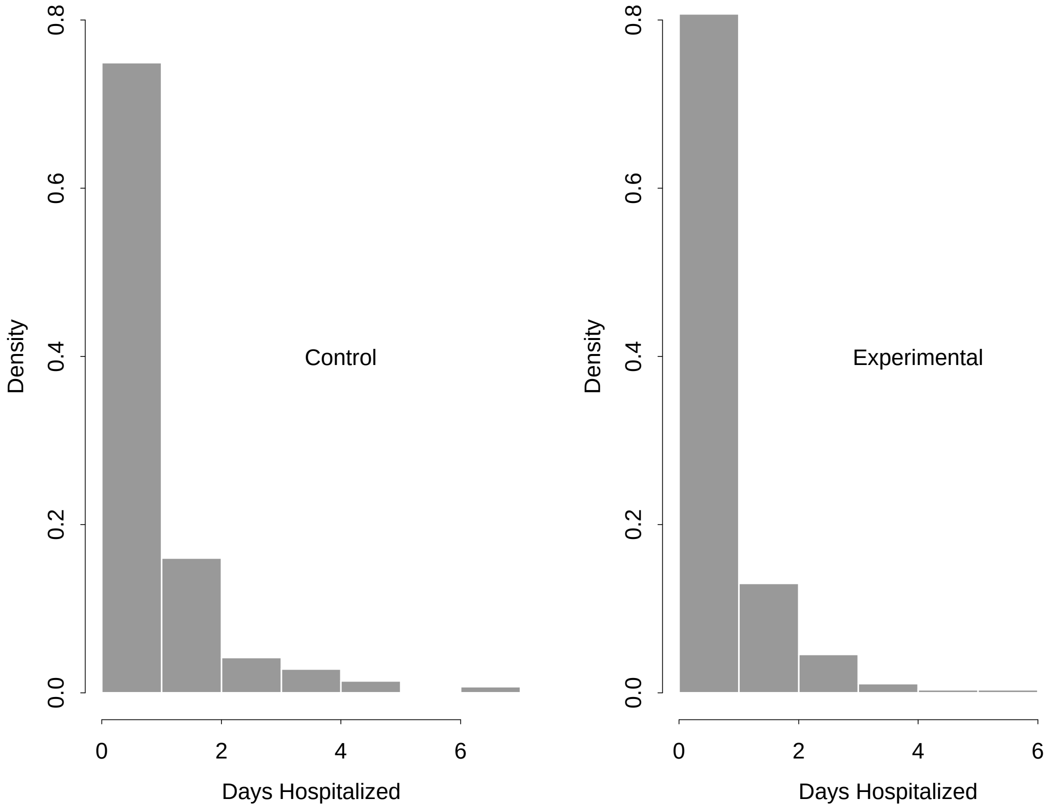

Figure 3.1. Histograms of control and treatment numbers of hospitalizations.

Initial Poisson Modeling

`S+` code to make the histograms:

```
> motif( )                          # to open a graphics window

> par( mfrow = c( 1, 2 ) )          # to plot two histograms

> hist( C, nclass = 8, probability = T,
    xlab = 'Days Hospitalized', ylab = 'Density',
    xlim = c( 0, 7 ), ylim = c( 0, 0.8 ) )

> text( 4, 0.4, 'Control' )

> hist( E, nclass = 8, probability = T,
    xlab = 'Days Hospitalized', ylab = 'Density',
    xlim = c( 0, 7 ), ylim = c( 0, 0.8 ) )

> text( 4, 0.4, 'Experimental' )
```

Instead of Model 0, Figure 3.1 suggests fitting a model of the form

$$
\begin{aligned}
(C_i|\lambda_C) &\sim \text{Poisson}\,(\lambda_C)\,, i = 1, \ldots, n_C, \\
\left(E_j|\lambda_E\right) &\sim \text{Poisson}\,(\lambda_E)\,, j = 1, \ldots, n_E,
\end{aligned}
\quad (21)
$$

perhaps with diffuse priors on λ_C and λ_E, and focusing on the **treatment effect** $\Delta = \lambda_E - \lambda_C$.

One way to do this is with **Markov Chain Monte Carlo (MCMC)** methods.

Simulation-Based Computation

Computation via conjugate analysis produces **closed-form results** (good) but is **limited in scope** to a fairly small set of models for which straightforward conjugate results are possible (bad).

This was a severe limitation for Bayesians for almost **250** years (from the 1750s to the 1980s).

Over the past 10 years the Bayesian community has "discovered" and developed an entirely new computing method, **Markov-Chain Monte Carlo (MCMC)** ("discovered" because the physicists first figured it out 47 years ago: Metropolis et al., 1953).

With a parameter vector θ and a data vector y, the idea is to **simulate** random draws from the posterior distribution $p(\theta|y)$, by constructing a **Markov chain**, (a) with the same state space as θ, (b) which is easy to simulate from, and (c) whose equilibrium distribution is $p(\theta|y)$ (if such a chain exists).

If you can do this, you can run the Markov chain for a long time, generating a huge sample from the posterior, and then use **simple descriptive summaries** (means, SDs, correlations, histograms or kernel density estimates) to extract any features of the posterior you want.

Gibbs Sampling

Two main types of MCMC have so far been developed: **Metropolis-Hastings (MH)** algorithms (e.g., Gilks et al., 1995), and **Gibbs sampling** (e.g., Gelfand and Smith, 1990). (Actually, Gibbs is a special case of MH, and M is a special case of H.) I will concentrate here on Gibbs sampling, which is somewhat less flexible but conceptually simpler.

Consider a problem in which the parameter θ is a vector $(\theta_1, \ldots, \theta_p)$. Given data $y = (y_1, \ldots, y_n)$, we are trying to compute the **full posterior** $p(\theta|y)$, various **marginal** distributions (Chapter 1, p. 31), e.g., $p(\theta_j|y)$, and **predictive** distributions such as $p(y_{n+1}|y)$.

With $\theta_{(j)}$ standing for θ with component j removed, the idea behind Gibbs is that the full posterior $p(\theta|y)$ will be characterized by its complete set of $\boxed{\textbf{full conditional}}$ distributions $p(\theta_j|y, \theta_{(j)})$, which will often be easier to write down in closed form than either $p(\theta|y)$ or $p(\theta_j|y)$.

This suggests an **iterative algorithm** of the following form: starting with an initial estimate $(\theta_1^{(0)}, \ldots, \theta_p^{(0)})$,

- Draw $\theta_1^{(1)}$ at random from $p(\theta_1|\, y, \theta_2^{(0)}, \ldots, \theta_p^{(0)})$;
- Draw $\theta_2^{(1)}$ at random from $p(\theta_2|\, y, \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_p^{(0)})$; and so on down to
- Draw $\theta_p^{(1)}$ at random from $p(\theta_p|\, y, \theta_1^{(1)}, \ldots, \theta_{p-1}^{(1)})$.

This constitutes one iteration of the Gibbs sampler. **Repeat until convergence** to the equilibrium distribution, and then repeat M times after that (for some large M) to **monitor** the posterior features of interest.

Practical Issues

From a given set of initial values you will probably want to discard the first B iterations—the **burn-in**—where B is chosen large enough that equilibrium has been reached.

Implementation Details. (1) How do you figure out the **full conditionals**, and how do you sample from them? (2) What should you use for **initial values**? (3) How **large** should B and M be? (4) More generally, how do you know when the chain has reached **equilibrium**?

Questions (3–4) fall under the heading of **MCMC diagnostics**, which will be covered in the second tutorial (question (2) is closely related to choice of B).

As for question (1), until recently when you wanted to make MCMC calculations you had to program up, e.g., a **Gibbs sampler from scratch**, using various tricks that were particular to your specific problem.

In the last 5 years a prototype version of a fairly general-purpose Gibbs sampling program, infelicitously named `BUGS`, has been produced by David Spiegelhalter, Wally Gilks, and others at the MRC Biostatistics Unit in Cambridge, UK (Spiegelhalter et al., 1997; available by **anonymous ftp** from `ftp.mrc-bsu.cam.ac.uk` or on the **web** at `http://www.mrc-bsu.cam.ac.uk`). (`BUGS` runs under Unix or DOS in a wide variety of hardware configurations, and a `Windows` version called `WinBUGS` is also available. I will concentrate on the non-`Windows` version here.)

BUGS and MLwiN

Another **computing environment** for Bayesian hierarchical modeling is MLwiN, a Windows-based package available at

http://www.ioe.ac.uk/mlwin/

(Bill Browne and I are the co-developers of the MCMC capabilities in this package; see **poster 9** on **Wednesday** evening.)

BUGS can fit a **wider variety** of models (not just HMs) than MLwiN, but if you do want to fit an HM MLwiN is typically **faster** (sometimes by quite a large margin).

IHGA Example, revisited. To implement Model 1 I created a file called poisson1.bug that looked like this:

```
model poisson1;

const

  n.C = 287, n.E = 285;

var

  lambda.C, lambda.E, C[ n.c ], E[ n.E ], effect;

data C in "poisson-C.dat", E in "poisson-E.dat";

inits in "poisson1.in";
```

Initial Poisson Modeling (continued)

```
{

  lambda.C ~ dgamma( 0.001, 0.001 );
  lambda.E ~ dgamma( 0.001, 0.001 );

  for ( i in 1:n.C ) {

    C[ i ] ~ dpois( lambda.C );

  }

  for ( j in 1:n.E ) {

    E[ j ] ~ dpois( lambda.E );

  }

  effect <- lambda.E - lambda.C;

}
```

`poisson1.in` initializes both λ_C and λ_E to 1.0; the $\Gamma(0.001, 0.001)$ priors for λ_C and λ_E are chosen to be **flat** in the region in which the **likelihood** is **appreciable**:

```
> sqrt( var( C ) / n.C )

[1] 0.07314114

> sqrt( var( E ) / n.E )

[1] 0.05972466

> c( mean( C ) - 3.0 * sqrt( var( C ) / n.C ),
    mean( C ) + 3.0 * sqrt( var( C ) / n.C ) )
```

Initial Poisson Modeling (continued)

```
[1] 0.7248275 1.1636743

> c( mean( E ) - 3.0 * sqrt( var( E ) / n.E ),
    mean( E ) + 3.0 * sqrt( var( E ) / n.E ) )

[1] 0.5892471 0.9475950

> lambda.grid <- seq( 0.01, 2.0, 0.01 )

> plot( lambda.grid, 0.001 * dgamma( lambda.grid, 0.001 ),
    type = 'l', xlab = 'Lambda', ylab = 'Density' )
```

The likelihood under the Gaussian model is **concentrated** for λ_C from about 0.7 to 1.2, and that for λ_E from about 0.6 to 1; you can see from the plot that across those ranges the $\Gamma(0.001, 0.001)$ prior is **essentially constant**.

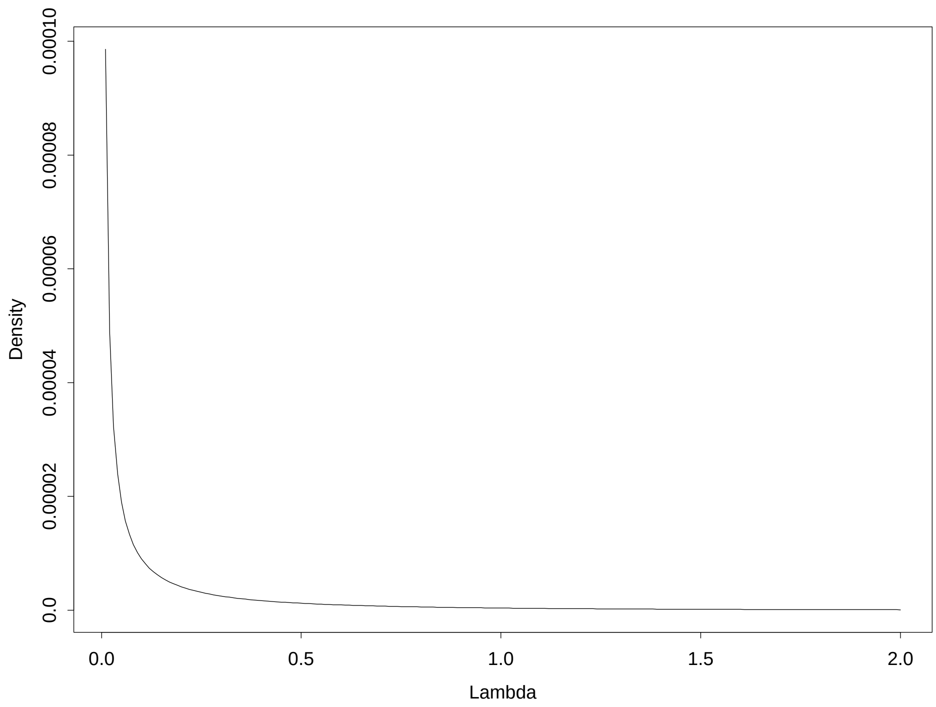

Figure 3.2. The $\Gamma(0.001, 0.001)$ distribution in the region in which the likelihoods for λ_C and λ_E are appreciable.

`WinBUGS` Implementation

(In the printed version this page contained **screen shots** from running the above model in `WinBUGS`.)

WinBUGS Implementation (continued)

(In the printed version this page contained **screen shots** from running the above model in WinBUGS.)

Initial Poisson Modeling (continued)

A burn-in of 2,000 and a monitoring run of 8,000 yields **good MCMC diagnostics** and permits a comparison between model 0 (Gaussian) and model 1 (Poisson), as in Table 3.3.

Table 3.3. Comparison of inferential conclusions from models 0 and 1.

$\boxed{\lambda_C}$ Model	Posterior Mean	Posterior SD	Central 95% Interval
Gaussian	0.944	0.0731	$(0.801, 1.09)$
Poisson	0.943	0.0577	$(0.832, 1.06)$

$\boxed{\lambda_E}$ Model	Posterior Mean	Posterior SD	Central 95% Interval
Gaussian	0.768	0.0597	$(0.651, 0.885)$
Poisson	0.769	0.0521	$(0.671, 0.875)$

$\boxed{\Delta = \lambda_E - \lambda_C}$ Model	Posterior Mean	Posterior SD	Central 95% Interval
Gaussian	-0.176	0.0944	$(-0.361, 0.009)$
Poisson	-0.174	0.0774	$(-0.325, -0.024)$

The two models produce **almost identical point estimates**, but the Poisson model leads to **sharper inferences** (e.g., the posterior SD for the treatment effect $\Delta = \lambda_E - \lambda_C$ is **22%** larger in model 0 than in model 1).

Additive vs. Multiplicative Effect

This is the same point we noticed with the NB10 data—when a location parameter is the only thing at issue, the Gaussian is a **conservative** modeling choice (the Poisson gains its "extra accuracy" from the variance and the mean being equal, which permits **second-moment** information to help in estimating the λ values along with the usual first-moment information).

Both the Gaussian and Poisson models so far implicitly assume that the treatment effect is **additive**:

$$E \stackrel{\text{st}}{=} C + \text{effect}, \tag{22}$$

where $\stackrel{\text{st}}{=}$ means *is stochastically equal to*; in other words, apart from random variation the effect of the IHGA is to **add or subtract a constant** to or from each person's underlying rate of hospitalization.

However, since the outcome variable is non-negative, it is plausible that a **better model** for the data is

$$E \stackrel{\text{st}}{=} (1 + \text{effect})\, C. \tag{23}$$

Additive vs. Multiplicative Effect

Here the treatment effect is **multiplicative**—in other words, apart from random variation the effect of the IHGA is to multiply each person's underlying rate of hospitalization by a constant above or below 1.

A **qqplot** of the control and experimental outcome values can be helpful in choosing between additive and multiplicative models:

```
> CEqq <- qqplot( C, E, plot = F )

> table( CEqq$y, CEqq$x )

                 Interpolated C values
       0 0.965   1 1.5   2 2.82 3 3.91 4 4.96 5 6.99 7
  0 137      1   9   0   0    0 0    0 0    0 0    0 0
  1   0      0  66   1  16    0 0    0 0    0 0    0 0
  2   0      0   0   0  29    1 7    0 0    0 0    0 0
E 3   0      0   0   0   0    0 4    1 7    1 0    0 0
  4   0      0   0   0   0    0 0    0 0    0 3    0 0
  5   0      0   0   0   0    0 0    0 0    0 0    1 0
  6   0      0   0   0   0    0 0    0 0    0 0    0 1

> symbols( c( 0, 0.964798, 1, 1, 1.5, 2, 2, 2.823944, 3, 3,
    3.908447, 4, 4.964813, 5, 6.985962, 7 ), c( rep( 0, 3 ),
    rep( 1, 3 ), rep( 2, 3 ), rep( 3, 4 ), 4, 5, 6 ),
    circles = c( 137, 1, 9, 66, 1, 16, 29, 1, 7, 4, 1, 7, 1,
    3, 1, 1 ), xlab = 'C', ylab = 'E' )
```

Additive vs. Multiplicative Effect

```
> abline( 0, 1 )                    # E = C (no effect)

> abline( 0, 0.793, lty = 2 )       # E = 0.816 C
                                    #   (multiplicative)

> abline( -0.174, 1, lty = 3 )      # E = C - 0.174 (additive)
```

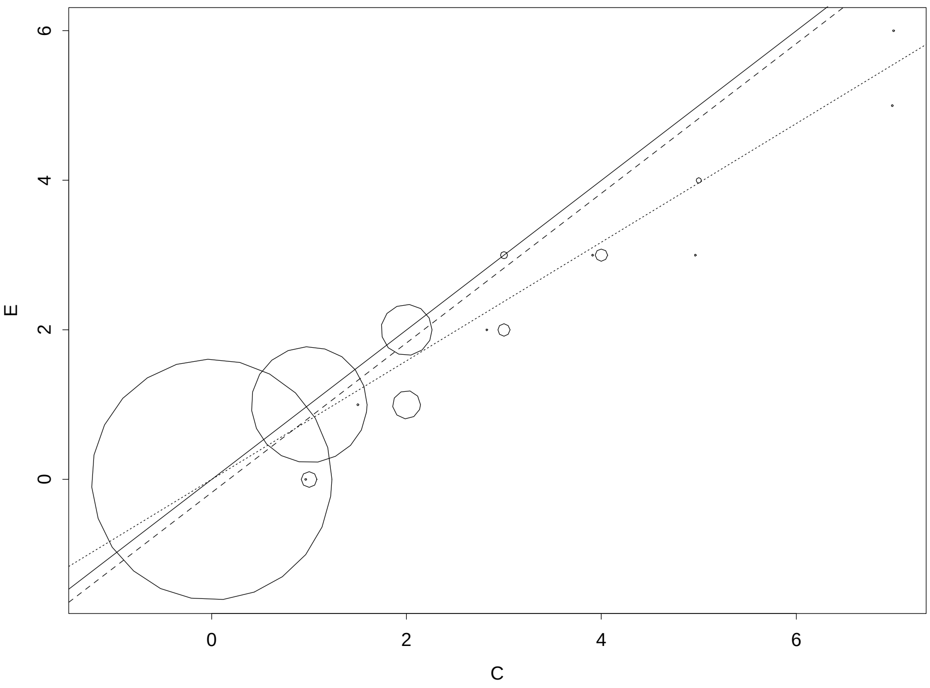

Figure 3.3. QQplot of E versus C values, with the radii of the plotted circles proportional to the number of observations at the indicated point. The solid line corresponds to no treatment effect, the small dotted line to the best-fitting multiplicative model ($E \stackrel{\text{st}}{=} 0.816\,C$), and the large dotted line to the best-fitting additive model ($E \stackrel{\text{st}}{=} C - 0.174$).

Because the Poisson model does not have separate parameters for location and scale, the **multiplicative** and **additive** formulations fit the data equally well, but the multiplicative model is easier to **generalize** in a meaningful way in this problem.

A Multiplicative Poisson Model

A simple way to write the multiplicative model is to re-express the data in the form of a **regression** of the outcome y on a **dummy variable** x which is 1 if the person was in the experimental group and 0 if he/she was in the control group:

i	1	2	$\cdots$	287	288	289	$\cdots$	572
x_i	0	0	$\cdots$	0	1	1	$\cdots$	1
y_i	1	0	$\cdots$	2	0	3	$\cdots$	1

Then for $i = 1, \ldots, n = 572$ the **multiplicative** model can be written

$$\begin{aligned} (y_i \,|\, \lambda_i) &\stackrel{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \gamma_0 + \gamma_1 x_i \\ (\gamma_0, \gamma_1) &\sim \text{diffuse} \end{aligned} \tag{24}$$

In this model the **control** people have

$$\log(\lambda_i) = \gamma_0 + \gamma_1(0) = \gamma_0, \quad \text{i.e.,} \quad \lambda_C = e^{\gamma_0}, \tag{25}$$

and the **experimental** people have

$$\begin{aligned} \log(\lambda_i) &= \gamma_0 + \gamma_1(1) = \gamma_0 + \gamma_1, \quad \text{i.e.,} \\ \lambda_E &= e^{\gamma_0 + \gamma_1} = e^{\gamma_0} e^{\gamma_1} = \lambda_C e^{\gamma_1}. \end{aligned} \tag{26}$$

Now you may remember from basic **Taylor series** that for γ_1 not too far from 0

$$e^{\gamma_1} \doteq 1 + \gamma_1, \tag{27}$$

A Multiplicative Poisson Model

so that finally (for γ_1 **fairly near 0**)

$$\lambda_E \doteq (1 + \gamma_1)\, \lambda_C, \tag{28}$$

which is a way of expressing equation (3) in Poisson language.

Fitting this model in BUGS is easy:

```
model poisson2;

const

  n = 572;

var

  gamma.0, gamma.1, lambda[ n ], x[ n ], y[ n ], lambda.C,
  lambda.E, mult.effect;

data x in "poisson-x.dat", y in "poisson-y.dat";
inits in "poisson2.in";

{
  gamma.0 ~ dnorm( 0.0, 1.0E-4 );    # flat priors for
  gamma.1 ~ dnorm( 0.0, 1.0E-4 );    # gamma.0 and gamma.1

  for ( i in 1:n ) {

    log( lambda[ i ] ) <- gamma.0 + gamma.1 * x[ i ];
    y[ i ] ~ dpois( lambda[ i ] );

  }

  lambda.C <- exp( gamma.0 );
  lambda.E <- exp( gamma.0 + gamma.1 );
  mult.effect <- exp( gamma.1 );
}
```

WinBUGS Implementation

(In the printed version this page contained **screen shots** from running the above model in WinBUGS.)

`WinBUGS` Implementation (continued)

(In the printed version this page contained **screen shots** from running the above model in `WinBUGS`.)

Additive versus Multiplicative Fit

A burn-in of 2,000 and a monitoring run of 8,000 again yields **good MCMC diagnostics** and permits a comparison between the additive and multiplicative Poisson models, as in Table 3.4.

Table 3.4. Comparison of inferential conclusions from the additive and multiplicative Poisson models.

$\boxed{\lambda_C}$ Model	Posterior Mean	Posterior SD	Central 95% Interval
additive	0.943	0.0577	$(0.832, 1.06)$
multiplicative	0.945	0.0574	$(0.837, 1.06)$

$\boxed{\lambda_E}$ Model	Posterior Mean	Posterior SD	Central 95% Interval
additive	0.769	0.0521	$(0.671, 0.875)$
multiplicative	0.768	0.0518	$(0.671, 0.872)$

$\boxed{\text{effect}}$ Model	Posterior Mean	Posterior SD	Central 95% Interval
additive	-0.174	0.0774	$(-0.325, -0.024)$
multiplicative	-0.184	0.0743	$(-0.324, -0.033)$

With this model it is as if the experimental people's average underlying rates of hospitalization have been **multiplied by 0.82**, give or take about 0.07.

The additive and multiplicative effects are **similar**, because the control mean was close to 1.

Extra-Poisson Variability

However, none of this has verified that the **Poisson model is reasonable** for these data—the histograms show that the Gaussian model is clearly unreasonable, but the diagnostic plots in `WinBUGS` only check on the adequacy of the **MCMC** sampling, not the model.

In fact we had a good clue that the data are **not** Poisson back on p. 2: the Poisson(λ) distribution has mean λ and also variance λ—in other words, the **variance-to-mean-ratio** (VTMR) for the Poisson is 1. But

```
> var( C ) / mean( C )
[1] 1.62599
> var( E ) / mean( E )
[1] 1.322979
```

i.e., the data exhibit **extra-Poisson variability** (VTMR > 1). This actually makes good sense if you think about it, as follows.

The Poisson model assumes that everybody in the control group has the **same underlying rate** λ_C of hospitalization, and similarly everybody in the experimental group has the **same rate** λ_E.

Unobserved Predictor Variables

In fact it is far more reasonable to think that each person has his/her **own** underlying rate of hospitalization that depends on **baseline health status**, **age**, and various other things.

Now Hendriksen **forgot to measure** (or at least to report on) these other variables (he may have hoped that the randomization would balance them between C and E)—the only predictor we have is x, the experimental status dummy variable—so the best we can do is to lump all of these other **unobserved** variables together into a kind of **"error" term** e.

This amounts to **expanding** the second Poisson model (24) above: for $i = 1, \ldots, n = 572$ the new model is

$$\begin{aligned} (y_i \,|\, \lambda_i) & \stackrel{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) & = \gamma_0 + \gamma_1 x_i + e_i \\ e_i & \stackrel{\text{IID}}{\sim} N\left(0, \sigma_e^2\right) \\ (\gamma_0, \gamma_1) & \sim \text{diffuse.} \end{aligned} \tag{29}$$

Random Effects Poisson Regression

The **Gaussian** choice for the error distribution is **conventional**, not dictated by the science of the problem (although if there were a lot of unobserved predictors hidden inside the e_i their weighted sum would be close to normal by the **Central Limit Theorem**).

Model (9) is an expansion of the earlier model (4) because you can obtain model (4) from (9) by setting $\sigma_e^2 = 0$, whereas with (9) we are letting it vary and learning about it from the data.

The addition of the $\boxed{\textbf{random effects}}$ e_i to the model is one way to address the extra-Poisson variability: this model would be called a **lognormal mixture of Poisson distributions** (or a **random effects Poisson regression** (REPR) model) because it is as if each person's λ is drawn from a lognormal distribution and then his/her number of hospitalizations y is drawn from a Poisson distribution with his/her λ, and this mixing process will make the variance of y **bigger than its mean**.

`WinBUGS` Implementation

The new `BUGS` model is

```
{

  gamma.0 ~ dnorm( 0.0, 1.0E-4 )
  gamma.1 ~ dnorm( 0.0, 1.0E-4 )
  tau.e ~ dgamma( 0.001, 0.001 )

  for ( i in 1:n ) {

    e[ i ] ~ dnorm( 0.0, tau.e )
    log( lambda[ i ] ) <- gamma.0 + gamma.1 * x[ i ] +
      e[ i ]
    y[ i ] ~ dpois( lambda[ i ] )

  }

  lambda.C <- exp( gamma.0 )
  lambda.E <- exp( gamma.0 + gamma.1 )
  mult.effect <- exp( gamma.1 )
  sigma.e <- 1.0 / sqrt( tau.e )

}
```

I again use a **diffuse** $\Gamma(\epsilon, \epsilon)$ prior (with $\epsilon = 0.001$) for the **precision** τ_e of the random effects.

WinBUGS Implementation

(In the printed version this page contained **screen shots** from running the above model in WinBUGS.)

Sensitivity to Initial Values

$\boxed{\textbf{Warning}}$ WinBUGS can fail, particularly in random-effects models, when you give it initial values that are not very close to the final posterior means; an example is the REPR model (9) on the IHGA data with the **"generic"** starting values $(\gamma_0, \gamma_1, \tau_e) = (0.0, 0.0, 1.0)$.

Two ways out in WinBUGS: **trial and error**, or a calculation (see below).

NB MLwiN does not have this problem—it gets its starting values from **maximum likelihood** (the mode of the likelihood function is often a decent approximation to the mean or mode of the posterior).

Technical note. To get a decent starting value for τ_e in model (9) you can calculate as follows: renaming the random effects η_i to avoid confusion with the number e, (1) $V(y_i) = V[E(y_i \,|\eta_i)] + E[V(y_i \,|\eta_i)]$, where (2) $(y_i \,|\eta_i) \sim$ Poisson $\left(e^{\gamma_0+\gamma_1 x_i+\eta_i}\right)$, so $E(y_i \,|\eta_i) = V(y_i \,|\eta_i) = e^{\gamma_0+\gamma_1 x_i+\eta_i}$. Then (3) $V[E(y_i \,|\eta_i)] = V\left(e^{\gamma_0+\gamma_1 x_i+\eta_i}\right) = e^{2(\gamma_0+\gamma_1 x_i)} V(e^{\eta_i})$ and $E[V(y_i \,|\eta_i)] = E\left(e^{\gamma_0+\gamma_1 x_i+\eta_i}\right) = e^{\gamma_0+\gamma_1 x_i} E(e^{\eta_i})$. Now (4) e^{η_i} is lognormal with mean 0 and variance σ_e^2 on the log scale, so $E(e^{\eta_i}) = e^{\frac{1}{2}\sigma_e^2}$ and $V(e^{\eta_i}) = e^{\sigma_e^2}\left(e^{\sigma_e^2} - 1\right)$, yielding finally $V(y_i) = e^{2(\gamma_0+\gamma_1 x_i)+\frac{1}{2}\sigma_e^2} + e^{\gamma_0+\gamma_1 x_i+\sigma_e^2}\left(e^{\sigma_e^2} - 1\right)$. (5) Plugging in $x_i = 0$ for the C group, whose sample variance is 1.54, and using the value $\gamma_0 = -0.29$ from runs with previous models, gives an equation for σ_e^2 that can be solved numerically, yielding $\sigma_e^2 \doteq 0.5$ and $\tau_e \doteq 2$.

`WinBUGS` Implementation (continued)

(In the printed version this page contained **screen shots** from running the above model in `WinBUGS`.)

WinBUGS Implementation (continued)

(In the printed version this page contained **screen shots** from running the above model in WinBUGS.)

REPR Model Results

(Another way to achieve a similar goal would be to fit different **negative binomial** distributions to the observed counts in the C and E groups—the negative binomial is a **gamma mixture of Poissons**, and the gamma and lognormal distributions often fit long-tailed data about equally well, so you would not be surprised to find that the two approaches give **similar results**.)

Table 4.5. Comparison of inferential conclusions about the multiplicative effect parameter e^{γ_1} from the fixed- and random-effects Poisson regression models.

Model	Posterior Mean	Posterior SD	Central 95% Interval
Poisson	0.816	0.0735	(0.683, 0.969)
REPR	0.830	0.0921	(0.665, 1.02)

Table 4.5 compares the REPR model inferential results with those from model (4), which could also be called a **fixed-effects Poisson regression model**.

The "error" SD σ_e has posterior mean **0.68**, give or take about 0.07 (on the $\log(\lambda)$ scale), corresponding to substantial extra-Poisson variability, which translates into **increased uncertainty** about the multiplicative effect parameter e^{γ_1}.

I will argue later that the REPR model **fits the data well**, so the conclusion I would publish from these data is that IHGA reduces the average number of hospitalizations per two years by about $100\,(1 - 0.083)\% = \boxed{\mathbf{17\%}}$ give or take about 9% (ironically this conclusion is similar to that from the Gaussian model, but this is coincidence).

3.2 Predictive Diagnostics

Model selection: A Bayesian perspective. An interest in diagnostics implies a willingness to evaluate and criticize a statistical model, i.e., to discover what (if anything) is deficient about the current "working" model and (if possible) to find a better model.

Thus the real topic here is **model selection**.

This is an active area of Bayesian research, both for hierarchical and other models (e.g., Leamer 1978, Spiegelhalter and Smith 1982, McCulloch and Rossi 1992, Lindley 1993, Key et al. 1998). It is clearly a confusing topic, to judge from the volume of Bayesian papers that descend into **ad-hockery** (e.g., Aitkin 1991, O'Hagan 1995, Gelman et al. 1996, Berger and Pericci 1996, Bayarri and Berger 1998), which is nature's way of telling you as a Bayesian to think harder about **problem formulation**:

One of the nice things about Bayes is that there is *only one legitimate way forward* in any given situation involving a decision that needs to be made: **maximization of expected utility** based on probabilities updated with Bayes' Theorem.

Model Selection: A Decision Problem

In the case of model selection it would seem self-evident that *to choose a model you have to say to what purpose the model will be put*, for how else will you know whether your model is good enough?

Specifying this purpose demands a **decision-theoretic basis for model choice**. To take two examples,

(1) If you are going to choose which of several ways to behave in the future, then the model has to be *good enough to reliably aid you in choosing the best behavior*; or

(2) If you wish to make a scientific summary of what is known, then—remembering that the hallmark of good science is good prediction—the model has to be *good enough to make sufficiently accurate predictions of observable outcomes* (in which the dimensions along which accuracy is to be monitored are driven by what is scientifically relevant).

Bayesian Decision Theory

Axiomatic approaches to rational decision-making date back to Ramsay (1926), with von Neumann and Morgenstern (1944) and Savage (1954) also making major contributions. The ingredients of a general decision problem (e.g., Bernardo and Smith, 1994) include

- A set $\{a_i, i \in I\}$ of available **actions**, one of which You will choose;

- For each action a_i, a set $\{E_j, j \in J\}$ of **uncertain outcomes** describing what will happen if You choose action a_i;

- A set $\{c_j, j \in J\}$ of **consequences** corresponding to the outcomes $\{E_j, j \in J\}$; and

- A **preference relation** $\leq$, expressing Your preferences between pairs of available actions ($a_1 \leq a_2$ means "a_1 is not preferred by You to a_2"). Define $a_1 \sim a_2$ ("a_1 and a_2 are *equivalent*" to You) iff $a_1 \leq a_2$ and $a_2 \leq a_1$.

Within this framework you have to make further assumptions—the **coherence** axioms—to ensure that your actions are internally consistent (see Bernardo and Smith 1994 for the formalism). These (reasonable) assumptions imply the existence and uniqueness of a (personal) probability $P_{B:\text{You}}$ (abbreviated P), mapping from outcomes E to [0,1] and corresponding to the judgements in Your definition of $\leq$, and a **utility function** U_{You} (abbreviated U; large values preferred, say), mapping from consequences c to R and quantifying Your preferences.

Utility-Based Variable Selection

Draper and Fouskakis (2000) give one example of this theory in action, demonstrating that variable selection in generalized linear models should often be governed by the principle that the final model should only contain variables that predict well enough **given how much they cost to collect** (Fig. 3.2, which compares $2^{14} = 16,384$ models).

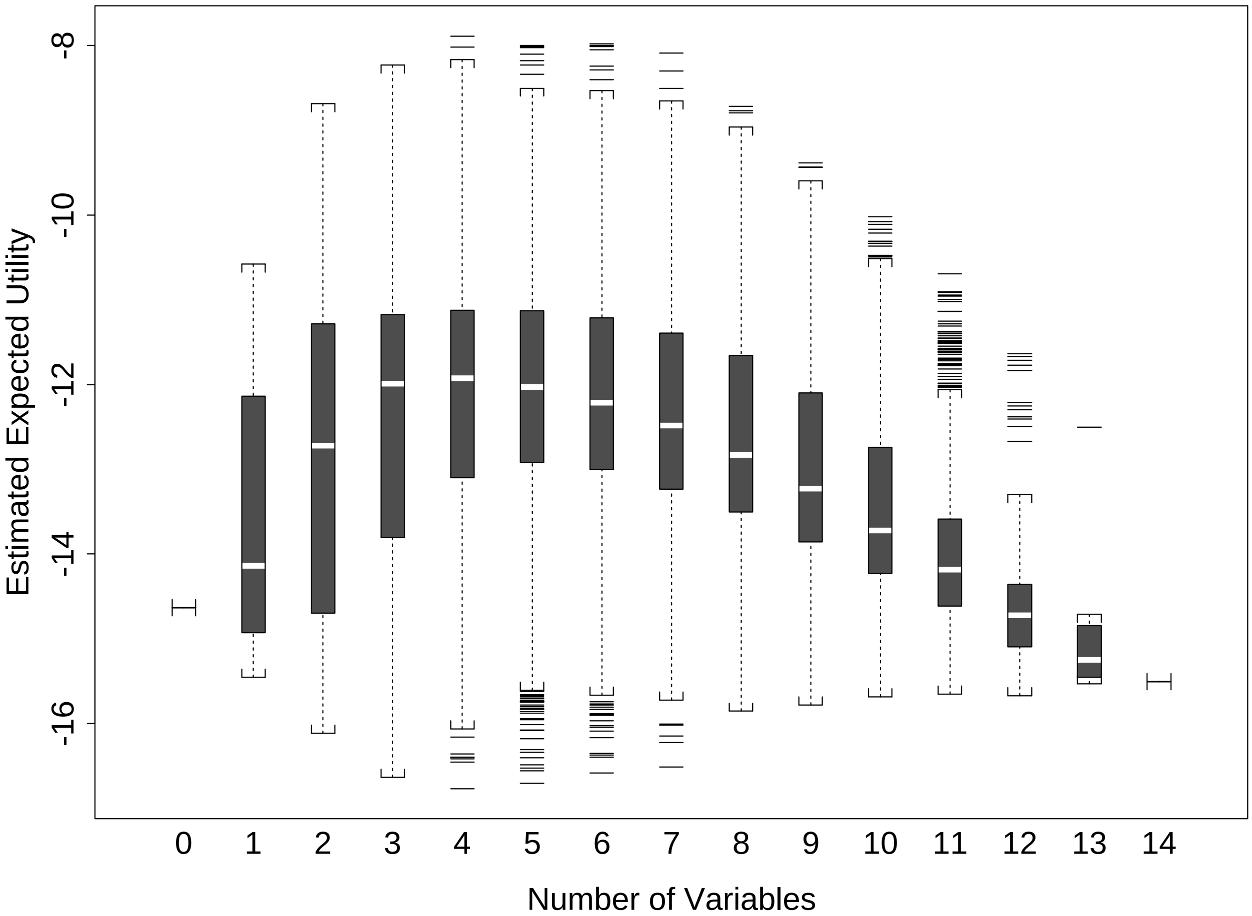

Fig. 3.4. *Estimated expected utility as a function of number of predictor variables, in a problem involving the construction of a cost-effective scale to measure the sickness at hospital admission of elderly pneumonia patients. The best models only have 4–6 sickness indicators out of the 14 possible predictors.*

Choosing the Utility Function

Any reasonable utility function in the admission sickness problem will have two components, one quantifying **data collection costs** associated with the construction of a given sickness scale, the other rewarding and penalizing the scale's **predictive successes and failures**.

This requires an intimate knowledge of the **real-world consequences** of correct choices and mistakes, a level of knowledge that is always desirable but costly in time and money to acquire.

In practice people often try to get away with a more **generic** utility function, e.g., one that simply rewards predictive accuracy. It is a mistake, however, to *use the data twice* in measuring this sort of thing (once to make predictions, once to see how good they are).

Out-of-sample predictive validation (e.g., Gelfand et al. 1992) solves this problem: e.g., successively remove each observation y_i one at a time, construct the predictive distribution for y_j based on y_{-j} (the data vector with y_j removed), and see where y_j falls in this distribution.

Log scoring rule

This motivates the **log scoring rule** (e.g., Bernardo and Smith 1994): with n data values y_j and when choosing among k models $M_i, i \in I$, find that model M_i which maximizes

$$\frac{1}{n} \sum_{j=1}^{n} \log p(y_j | M_i, y_{-j}). \tag{30}$$

It is also revealing to compute **predictive z–scores**, for observation j under model i:

$$z_{ij} = \frac{y_j - E(y_j | M_i, y_{-j})}{\sqrt{V(y_j | M_i, y_{-j})}}. \tag{31}$$

A serious Bayesian modeling problem: Cromwell's Rule. Lindley (1972) reminds us of what he calls *Cromwell's rule*: Anything with zero prior probability has to have zero posterior probability, no matter how the data come out.

This gives Bayesians fits when applied to model selection: if you pick an initial model before seeing the data, thus giving zero prior probability to huge regions of model space, and the data make you wish that you had enlarged the initial model, **what can you do without cheating** (using the data twice)?

3CV: Modeling, Validation, Projection

One possibility: **Bayesian nonparametrics** (if your prior on model space spans all (scientifically relevant) possibilities then Cromwell's Rule disappears).

Another possibility (3CV: Draper 1998): Taking the cross-validation idea one step further,

(1) Divide data at random into three subsets S_i.

(2) Fit tentative {likelihood + prior} to S_1. **Expand** initial model in all feasible ways suggested by data exploration using S_1. **Iterate** until you're happy.

(3) Use final model (fit to S_1) from (2) to create predictive distributions for all data points in S_2. Compare actual outcomes with these distributions, checking for **predictive calibration**. Go back to (2), change likelihood as necessary, **retune prior** as necessary, to get good calibration. Iterate until you're happy.

(4) Announce **final model** (fit to S_1 and S_2) from (3), and report predictive calibration of this model on data points in S_3 as indication of how well it would perform with new data.

Hendriksen data revisited

Experimental Status	Model	Validation	Projection	Total n
Treatment	143	71	71	285
Control	143	72	72	287
Total n	286	143	143	572

```
log.score.summary <- function( treatment, control, n3cv.T,
  n3cv.C, S, N ) {

  n.T <- length( treatment )
  n.C <- length( control )

  log.score1.V <- rep( 0, S )
  log.score2.V <- rep( 0, S )
  log.score3.V <- rep( 0, S )
  log.score4.V <- rep( 0, S )

  z.summary1.V <- matrix( 0, S, 6 )
  z.summary2.V <- matrix( 0, S, 6 )
  z.summary3.V <- matrix( 0, S, 6 )
  z.summary4.V <- matrix( 0, S, 6 )

  for ( i in 1:S ) {

#     general stuff for all models

    rand.T <- sample( 1: n.T )
    y.TM <- treatment[ rand.T[ 1:n3cv.T[1] ] ]
    y.TV <- treatment[ rand.T[ ( n3cv.T[1] + 1 ):
      ( n3cv.T[1] + n3cv.T[2] ) ] ]
    y.TP <- treatment[ rand.T[ ( n3cv.T[1] + n3cv.T[2]
      + 1 ): n.T ] ]

    rand.C <- sample( 1: n.C )
    y.CM <- control[ rand.C[ 1:n3cv.C[1] ] ]
    y.CV <- control[ rand.C[ ( n3cv.C[1] + 1 ):
      ( n3cv.C[1] + n3cv.C[2] ) ] ]
    y.CP <- control[ rand.C[ ( n3cv.C[1] + n3cv.C[2]
      + 1 ): n.C ] ]
```

S+ Code (continued)

```
    y.M <- c( y.CM, y.TM )
    y.V <- c( y.CV, y.TV )
    y.P <- c( y.CP, y.TP )

    n.M <- length( y.M )
    n.V <- length( y.V )
    n.P <- length( y.P )

    n.TM <- length( y.TM )
    n.CM <- length( y.CM )
    n.TV <- length( y.TV )
    n.CV <- length( y.CV )
    n.TP <- length( y.TP )
    n.CP <- length( y.CP )

#     stuff specific to model 1

    ybar.TM <- mean( y.TM )
    s.TM <- sqrt( var( y.TM ) )

    ybar.CM <- mean( y.CM )
    s.CM <- sqrt( var( y.CM ) )

    log.score1.TV <- sum( log( dnorm( y.TV, ybar.TM, s.TM ) ) )
    log.score1.CV <- sum( log( dnorm( y.CV, ybar.CM, s.CM ) ) )

    log.score1.V[i] <- ( log.score1.TV + log.score1.CV ) / n.V

    z.scores1.TV <- ( y.TV - ybar.TM ) / s.TM
    z.scores1.CV <- ( y.CV - ybar.CM ) / s.CM

    z.scores1.V <- c( z.scores1.TV, z.scores1.CV )

    qq1 <- qqnorm( z.scores1.V, plot = F )

    z.summary1.V[i,] <- c( mean( z.scores1.V ),
      var( z.scores1.V ), sum( abs( z.scores1.V ) > 1.959964 )
      / n.V, sum( abs( z.scores1.V ) > 2.575829 ) / n.V,
      sum( abs( z.scores1.V ) > 3.290527 ) / n.V,
      cor( qq1$x, qq1$y ) )
```

S+ Code (continued)

```
#     stuff specific to model 2

    r.M <- sum( y.M )

    log.score2.V[i] <- sum( log( dpg( y.V, r.M, n.M, 1 ) ) ) /
      n.V

    pred.mean2 <- r.M / n.M

    pred.SD2 <- sqrt( r.M * ( 1.0 + 1.0 / n.M ) / n.M )

    z.scores2.V <- ( y.V - pred.mean2 ) / pred.SD2

    qq2 <- qqnorm( z.scores2.V, plot = F )

    z.summary2.V[i,] <- c( mean( z.scores2.V ),
      var( z.scores2.V ), sum( abs( z.scores2.V ) > 1.959964 )
      / n.V, sum( abs( z.scores2.V ) > 2.575829 ) / n.V,
      sum( abs( z.scores2.V ) > 3.290527 ) / n.V,
      cor( qq2$x, qq2$y ) )

#     stuff specific to model 3

    r.TM <- sum( y.TM )
    log.score3.TV <- sum( log( dpg( y.TV, r.TM, n.TM, 1 ) ) )

    r.CM <- sum( y.CM )
    log.score3.CV <- sum( log( dpg( y.CV, r.CM, n.CM, 1 ) ) )

    log.score3.V[i] <- ( log.score3.TV + log.score3.CV ) / n.V

    pred.meanT3 <- r.TM / n.TM
    pred.SDT3 <- sqrt( r.TM * ( 1.0 + 1.0 / n.TM ) / n.TM )
    z.scores3.TV <- ( y.TV - pred.meanT3 ) / pred.SDT3

    pred.meanC3 <- r.CM / n.CM
    pred.SDC3 <- sqrt( r.CM * ( 1.0 + 1.0 / n.CM ) / n.CM )
    z.scores3.CV <- ( y.CV - pred.meanC3 ) / pred.SDC3

    z.scores3.V <- c( z.scores3.TV, z.scores3.CV )

    qq3 <- qqnorm( z.scores3.V, plot = F )
```

S+ Code (continued)

```
    z.summary3.V[i,] <- c( mean( z.scores3.V ),
      var( z.scores3.V ), sum( abs( z.scores3.V ) > 1.959964 )
      / n.V, sum( abs( z.scores3.V ) > 2.575829 ) / n.V,
      sum( abs( z.scores3.V ) > 3.290527 ) / n.V,
      cor( qq3$x, qq3$y ) )

#     stuff specific to model 4

    write( c( rep( 0, n.CM ), rep( 1, n.TM ) ), "poisson-x.dat",
      ncol = 20, append = F )
    write( y.M, "poisson-y.dat", ncol = 20, append = F )

    junk <- unix( "backbugs poisson4.cmd" )

    temp <- matrix( scan( "bugs.out" ), 2 * N, 2, byrow = T )

    pred.dist.x0.data <- temp[ 1:N, 2 ]
    top.0 <- max( pred.dist.x0.data )
    pred.dist.x0 <- hist( pred.dist.x0.data, breaks = ( -0.5 +
      ( 0:( top.0 + 1 ) ) ), plot = F )$counts / N
    pred.dist.x0 <- pred.dist.x0 + min( pred.dist.x0[
      pred.dist.x0 > 0.0 ] )
    pred.dist.x0 <- pred.dist.x0 / sum( pred.dist.x0 )

    pred.dist.x1.data <- temp[ ( N + 1 ):( 2 * N ), 2 ]
    top.1 <- max( pred.dist.x1.data )
    pred.dist.x1 <- hist( pred.dist.x1.data, breaks = ( -0.5 +
      ( 0:( top.1 + 1 ) ) ), plot = F )$counts / N
    pred.dist.x1 <- pred.dist.x1 + min( pred.dist.x1[
      pred.dist.x1 > 0.0 ] )
    pred.dist.x1 <- pred.dist.x1 / sum( pred.dist.x1 )

    log.score4.CV <- sum( log( pred.dist.x0[ 1 + y.CV ] ) )
    log.score4.TV <- sum( log( pred.dist.x1[ 1 + y.TV ] ) )
    log.score4.V[i] <- ( log.score4.CV + log.score4.TV ) / n.V

    pred.meanC4 <- mean( pred.dist.x0.data )
    pred.SDC4 <- sqrt( var( pred.dist.x0.data ) )
    z.scores4.CV <- ( y.CV - pred.meanC4 ) / pred.SDC4

    pred.meanT4 <- mean( pred.dist.x1.data )
    pred.SDT4 <- sqrt( var( pred.dist.x1.data ) )
    z.scores4.TV <- ( y.TV - pred.meanT4 ) / pred.SDT4
```

S+ Code (continued)

```
    z.scores4.V <- c( z.scores4.TV, z.scores4.CV )

    qq4 <- qqnorm( z.scores4.V, plot = F )

    z.summary4.V[i,] <- c( mean( z.scores4.V ),
      var( z.scores4.V ), sum( abs( z.scores4.V ) > 1.959964 )
      / n.V, sum( abs( z.scores4.V ) > 2.575829 ) / n.V,
      sum( abs( z.scores4.V ) > 3.290527 ) / n.V,
      cor( qq4$x, qq4$y ) )

    write( signif( c( i, log.score1.V[i], z.summary1.V[i,],
      log.score2.V[i], z.summary2.V[i,], log.score3.V[i],
      z.summary3.V[i,], log.score4.V[i], z.summary4.V[i,] ),
      7 ), "poisson4.raw", ncol = 29, append = ( i > 1 ) )

  }

  return( cbind( log.score1.V, z.summary1.V, log.score2.V,
    z.summary2.V, log.score3.V, z.summary3.V, log.score4.V,
    z.summary4.V ) )

}

dpg <- function( y, alpha, beta, n ) {

  return( exp( alpha * log( beta ) + lgamma( alpha + y ) + y *
    log( n ) - lgamma( alpha ) - lgamma( y + 1 ) - ( alpha +
    y ) * log( beta + n ) ) )

}
```

BUGS code

```
compile( "poisson4.bug" )
update( 100 )
monitor( y.x0.new )          # This is
monitor( y.x1.new )          # poisson4.cmd
update( 2500 )
q()

##########################

model poisson4;

const                        # And this is
                             # poisson4.bug
  n = 286;

var

  x[n], y[n], gamma.0, gamma.1, tau.e, x.bar, e[n],
  lambda[n], lambda.C, lambda.E, mult.effect, sd.e,
  e.x0.new, lambda.x0.new, y.x0.new, e.x1.new,
  lambda.x1.new, y.x1.new;

data x in "poisson-x.dat", y in "poisson-y.dat";
inits in "poisson4.in";

{

  gamma.0 ~ dnorm( 0.0, 0.0001 );
  gamma.1 ~ dnorm( 0.0, 0.0001 );
  tau.e ~ dgamma( 0.0001, 0.0001 );

  x.bar <- mean( x[] );

  for (i in 1:n) {

    e[i] ~ dnorm( 0.0, tau.e );
    log( lambda[i] ) <- gamma.0 + gamma.1 * ( x[i] - x.bar )
      + e[i];
    y[i] ~ dpois( lambda[i] );

  }
```

BUGS code (continued)

```
  lambda.C <- exp( gamma.0 );
  lambda.E <- exp( gamma.0 + gamma.1 );
  mult.effect <- exp( gamma.1 );
  sd.e <- 1.0 / sqrt( tau.e );

  e.x0.new ~ dnorm( 0.0, tau.e );
  log( lambda.x0.new ) <- gamma.0 + gamma.1 * ( 0.0 - x.bar )
      + e.x0.new;
  y.x0.new ~ dpois( lambda.x0.new );

  e.x1.new ~ dnorm( 0.0, tau.e );
  log( lambda.x1.new ) <- gamma.0 + gamma.1 * ( 1.0 - x.bar )
      + e.x1.new;
  y.x1.new ~ dpois( lambda.x1.new );

}
```

Results

simulation SEs in parenthesis

Model 1: Gaussian

S	log score	mean z-score	var z-score	% > z(.05)	% > z(.01)	% > z(.001)	qqnorm corr
5000	-1.552	0.005086	1.050	0.05637	0.02647	0.01358	0.8743
	(0.001)	(0.001469)	(0.004)	(0.00027)	(0.00024)	(0.00015)	(0.0003)

This is OK on mean and variance grounds but is missing something in the tail(s) from the unusual z-values.

Model 2: Poisson with common lambda

S	log score	mean z-score	var z-score	% > z(.05)	% > z(.01)	% > z(.001)	qqnorm corr
5000	-1.316	0.002293	1.507	0.07700	0.03417	0.02951	0.8611
	(0.001)	(0.001755)	(0.004)	(0.00027)	(0.00021)	(0.00022)	(0.0003)

Better log score but MUCH worse var z-score and % unusual z-values.

Model 3: Poisson with different lambdas

S	log score	mean z-score	var z-score	% > z(.05)	% > z(.01)	% > z(.001)	qqnorm corr
5000	-1.314	0.004297	1.493	0.07438	0.04352	0.02318	0.8775
	(0.001)	(0.001750)	(0.004)	(0.00030)	(0.00028)	(0.00020)	(0.0003)

Evidently (from the z-score variances) the Poisson is underdispersed.

Model 4: random-effects Poisson regression

S	log score	mean z-score	var z-score	% > z(.05)	% > z(.01)	% > z(.001)	qqnorm corr
5000	-1.18	0.0013	1.02	.052	.011	.002	.942
	(0.001)	(0.0016)	(.004)	(.003)	(.002)	(.001)	(.001)

This model fits really well.

References

Bernardo JM, Smith AFM (1994). *Bayesian Theory*. New York: Wiley.

Bryk AS, Raudenbush SW (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. London: Sage.

Carlin BP, Louis TA (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Craig PS, Goldstein M, Seheult AH, Smith JA (1997). Constructing partial prior specifications for models of complex physical systems. *The Statistician*, **46**, forthcoming.

Draper D, Gaver D, Goel P, Greenhouse J, Hedges L, Morris C, Tucker J, Waternaux C (1993a). *Combining Information: Statistical Issues and Opportunities for Research*. Contemporary Statistics Series, No. 1. American Statistical Association, Alexandria VA.

Draper D, Hodges JS, Mallows CL, Pregibon D (1993b). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A*, **156**, 9–37.

Draper D (1995). Inference and hierarchical modeling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics*, **20**, 115–147, 233–239.

Draper D, Hodges JS, Mallows CL, Pregibon D (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A*, **156**, 9–37.

de Finetti B (1930). Funzione caratteristica di un fenomeno aleatorio. *Mem. Acad. Naz. Lincei*, **4**, 86–133.

de Finetti B (1964). Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability*, H. E. Kyburg, Jr., and H. E. Smokler, eds., New York: Wiley (1980), 93–158.

de Finetti B (1974/5). *Theory of Probability*, **1–2**. New York: Wiley.

Fisher RA (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, **222**, 309–368.

References (continued)

Fisher RA (1956). *Statistical Methods and Scientific Inference*. London: Oliver and Boyd.

Freedman D, Pisani R, Purves R, Adhikari A (1991). *Statistics*, second edition. New York: Norton.

Gelman A, Carlin JB, Stern HS, Rubin DB (1995). *Bayesian Data Analysis*. London: Chapman & Hall.

Hacking I (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.

Hedges LV, Olkin I (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.

Johnson NL, Kotz S (1970). *Distributions in statistics: Continuous univariate distributions*, **1**. New York: Wiley.

Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, **75**, 845–854.

Kadane JB, Wolfson LJ (1997). Experiences in elicitation. *The Statistician*, **46**, forthcoming.

Kahn K, Rubenstein L, Draper D, Kosecoff J, Rogers W, Keeler E, Brook R (1990). The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series. *Journal of the American Medical Association*, **264**, 1953–1955 (with editorial comment, 1995–1997).

Laplace PS (1774). Mémoire sur la probabilité des causes par les évenements. *Mémoires de l'Academie des Sciences de Paris*, **6**, 621–656. English translation in 1986 as "Memoir on the probability of the causes of events," with an introduction by SM Stigler, *Statistical Science*, **1**, 359–378.

Morris CN (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, **78**, 47–59.

References (continued)

Morris CN (1988). Determining the accuracy of Bayesian empirical Bayes estimators in the familiar exponential families. In *Proceedings of the Fourth Purdue Symposium on Statistical Decision Theory and Related Topics IV, part 1.*, SS Gupta, JO Berger, eds. New York: Springer-Verlag, 251–263.

O'Hagan A (1997). Eliciting expert beliefs in substantial practical applications. *The Statistician*, **46**, forthcoming.

Raudenbush SW (1984). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: A synthesis of findings from 19 experiments. *Journal of Educational Psychology*, **76**, 85–97.

Samaniego FJ, Reneau DM (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association*, **89**, 947–957.

Tierney L, Kadane JB (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.