One-Day Short Course on Bayesian Modeling, Inference and Prediction

7: Bayesian Model Specification

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
http://www.ams.ucsc.edu/∼draper

Sponsored by the Boston Chapter of the American Statistical Association

10 December 2004, 8am–5.30pm
Hotel@MIT, 20 Sidney Street, Cambridge MA

(the ideas in this section are joint with **Milovan Krnjajić** and represent work-in-progress on his Ph.D. dissertation)

© *2004 David Draper (all rights reserved)*

7.1 What is a Bayesian Model?

de Finetti (1970):

Bayesian model = joint predictive distribution

$$p(y) = p(y_1, \ldots, y_n) \tag{1}$$

for as-yet-unobserved **observables** $y = (y_1, \ldots, y_n)$.

Example 1: Data = **health outcomes** for all patients at one hospital with heart attack admission diagnosis.

Simplest possible: $y_i = 1$ if patient i **dies within 30 days of admission**, 0 otherwise.

de Finetti (1930): **in absence of any other information**, predictive uncertainty about y_i is **exchangeable**.

Representation theorem for binary data: if $(y_1, \ldots, y_n)$ part of **infinitely exchangeable sequence**, all **coherent** joint predictive distributions $p(y_1, \ldots, y_n)$ must have simple **hierarchical** form

$$\begin{aligned} \theta &\sim p(\theta) \\ (y_i|\theta) &\overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \end{aligned} \tag{2}$$

where $\theta = P(y_i = 1)$ = **limiting value of mean of** y_i in infinite sequence.

Model = Prior (Sometimes)

Mathematically $p(\theta)$ is **mixing distribution** in

$$p(y_1, \ldots, y_n) = \int_0^1 \theta^s (1-\theta)^{n-s} p(\theta)\, d\theta, \qquad (3)$$

where $s = \sum_{i=1}^n y_i$; **statistically**, $p(\theta)$ provides opportunity to quantify **prior information** about θ and combine with information in y.

Thus, in simplest situation, **Bayesian model specification** = choice of **scientifically appropriate prior distribution** $p(\theta)$.

Example 2 (elaborating Example 1): Now want to predict real-valued **sickness-at-admission score** instead of mortality.

Uncertainty about y_i still **exchangeable**; de Finetti's (1937) **representation theorem** for real-valued data: if $(y_1, \ldots, y_n)$ part of **infinitely exchangeable sequence**, all **coherent** joint predictive distributions $p(y_1, \ldots, y_n)$ must have (**no longer quite so simple**) hierarchical form

$$\begin{aligned} F &\sim p(F) \\ (y_i|F) &\overset{\text{IID}}{\sim} F, \end{aligned} \qquad (4)$$

where F = **limiting empirical cumulative distribution function** (CDF) of infinite sequence $(y_1, y_2, \ldots)$.

Bayesian Non- and Semi-Parametrics

Thus here Bayesian model specification = choosing **scientifically appropriate mixing (prior) distribution** $p(F)$ for F.

However, F is **infinite-dimensional parameter**; putting probability distribution on $\mathcal{D} = \{$all possible CDFs$\}$ is harder.

Specifying distributions on **function spaces** is task of Bayesian **non-parametric** (BNP), **semi-parametric** (BSP) modeling (e.g., Dey et al. 1998).

Example 3 (elaborating Example 2): In practice, in addition to **outcomes** y_i, **covariates** x_{ij} will typically be available.

For instance (Hendriksen et al. 1984), 572 elderly people **randomized**, 287 to **control** (C) group (standard care) and 285 to **treatment** (T) group (standard care plus **in-home geriatric assessment** (IHGA): **preventive medicine** in which each person's medical/social needs assessed, acted upon individually).

One important **outcome** was **number of hospitalizations** (in two years).

y_i^T, y_j^C = numbers of hospitalizations for **treatment** person i, **control** person j, respectively.

Suppose **treatment/control** (T/C) status is **only available covariate**.

Conditional Exchangeability

Unconditional judgment of exchangeability across all 572 outcomes **no longer automatically scientifically appropriate**.

Instead **design of experiment** compels (at least initially) judgment of **conditional exchangeability given T/C status** (e.g., de Finetti 1938, Draper et al. 1993), as in

$$\begin{array}{ccc} (F_T, F_C) & \sim & p(F_T, F_C) \\ (y_i^T | F_T, F_C) \overset{\text{IID}}{\sim} F_T & \Big| & (y_j^C | F_T, F_C) \overset{\text{IID}}{\sim} F_C \end{array} \tag{5}$$

It will later be necessary to decide if F_T, F_C are sufficiently **similar** that data are consistent with judgment of **unconditional** exchangeability (to see if treatment has **effect** or not).

This framework, in which (1) **covariates** specify **conditional exchangeability judgments**, (2) de Finetti's **representation theorem** reduces model specification task to placing appropriate prior distributions on CDFs, covers much of field of **statistical inference/prediction**; thus BNP/BSP modeling seem **crucial to entire Bayesian enterprise** over next 10–20 years.

Data-Analytic Model Specification

However, placing prior distributions on CDFs is **hard work**, we don't have much **experience** with it yet; in meantime, in parallel with efforts to **accumulate** experience, people will still do **parametric** modeling, we need **good tools** for specifying such models.

Standard (**data-analytic**) approach to model specification involves initial choice, for **structure** of model, of **standard parametric family**, followed by **modification** of initial choice—once data begin to arrive—if data suggest **deficiencies** in original specification.

This approach (e.g., Draper 1995) is **incoherent**: it uses data both to specify **prior distribution on structure space** and to **update** using **data-determined prior** (result will typically be **uncalibrated** (too narrow) predictive distributions for future data).

Dilemma is example of **Cromwell's Rule**: initial model choice placed 0 prior probability on large regions of model space; formally all such regions **must also have 0 posterior probability** even if data indicate **different prior on model space** would have been better.

Two Possible Solutions

• If use prior on F that places **non-zero probability on all Kullback-Leibler neighborhoods of all densities** (Walker et al. 2003; e.g., Pólya trees, Dirichlet process mixture priors, when chosen well), then BNP/BSP **directly avoids** Cromwell's Rule dilemma, at least for large n: as $n \to \infty$ posterior on F will **shrug off** any incorrect details of prior specification, will **fully adapt** to actual data-generating F (**NB** this assumes correct exchangeability judgments).

• **Three-way cross-validation** (3CV; Draper and Krnjajić 2004): taking usual cross-validation idea one step further,

(1) Divide data at random into *three* (non-overlapping) 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, 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.

With **large** n probably only need to do this **once**; with **small** and **moderate** n probably best to **repeat** (1–4) several times and **combine** results in some appropriate way (e.g., **model averaging**).

7.2 Model Selection
As a Decision Problem

Given method like 3CV which permits **hunting around in model space** without forfeiting calibration, **how do you know when to stop**?

It would seem self-evident that **to choose model you have to say to what purpose model will be put**, for how else will you know whether model is good enough?

Specifying this purpose demands
decision-theoretic basis for model choice
(e.g., Draper 1996; Key et al. 1998).

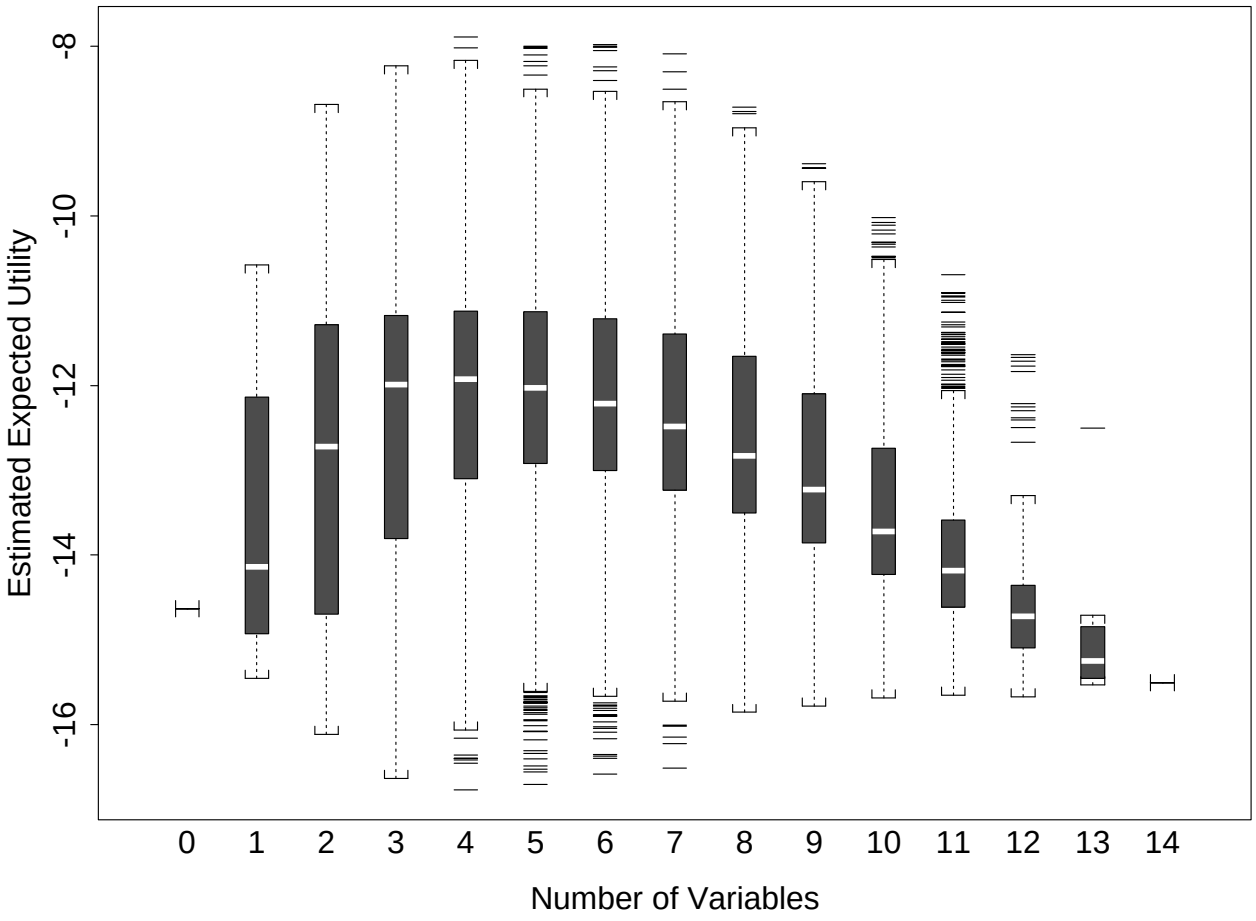
To take **two examples**,

(1) If you're going to choose which of several ways to behave in future, then model has to be **good enough to reliably aid in choosing best behavior** (see, e.g., Draper and Fouskakis example below); or

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

Utility-Based Variable Selection

Example 4: Draper and Fouskakis (2000, 2004) (also see Fouskakis and Draper 2002) give one example of decision-theoretic model choice in action, demonstrating that **variable selection in regression models** should often be governed by principle that final model should only contain variables that predict well enough **given how much they cost to collect** (see the figure below, which compares $2^{14} = \mathbf{16{,}384}$ models).



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

Choosing Utility Function

Any reasonable utility function in Example 4 will have two components, one quantifying **data collection costs** associated with construction of given sickness scale, other rewarding and penalizing scale's **predictive successes, failures**.

This requires intimate knowledge of **real-world consequences** of correct choices, mistakes—a level of understanding that's always desirable but is frequently **costly** in time, money to acquire.

Sometimes the main goal instead is **summary of scientific knowledge**, which suggests (as noted above) a utility function that rewards **predictive accuracy**.

On calibration grounds it's mistake, however, to **use data twice** in measuring this sort of thing (once to make predictions, again with same data 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_j one at a time, construct predictive distribution for y_j based on y_{-j} (data vector with y_j removed), see where y_j falls in this distribution.

Log Score as Utility

This motivates **log scoring rule** (e.g., Good 1950; Bernardo and Smith 1994): with n data values y_j, when choosing among k models $M_i, i \in I$, find that model M_i which maximizes

$$\frac{1}{n} \sum_{j=1}^{n} \log p(y_j | M_i, y_{-j}). \tag{6}$$

This can be given direct **decision-theoretic justification**: with utility function for model i

$$U(M_i|y) = \log p(y^* | M_i, y), \tag{7}$$

where y^* is **future data value**, expectation in MEU is over **uncertainty about** y^*; this expectation can be **estimated** (assuming exchangeability) by (6).

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

$$z_{ij} = \frac{y_j - E(y_j | M_i, y_{-j})}{\sqrt{V(y_j | M_i, y_{-j})}}. \tag{8}$$

For **good predictive calibration** of M_i, $\{z_{ij}, j = 1, \ldots, n\}$ should have **mean 0**, **standard deviation** (SD) **1**; often find instead that SD is **larger than 1** (predictive uncertainty bands **not wide enough**).

7.3 Approximating Log Score Utility

With **large data sets**, in situations in which **predictive distribution** has to be **estimated by MCMC**, log score expected utility (6) is **computationally expensive**; need **fast approximation** to it.

To see how this might be obtained, examine log score in **simplest possible model** M_0: for $i = 1, \ldots, n$,

$$\begin{aligned} \mu &\sim N\left(\mu_0, \sigma_\mu^2\right) \\ (Y_i|\mu) &\overset{\text{IID}}{\sim} N(\mu, \sigma^2) \end{aligned} \tag{9}$$

with σ known, take **highly diffuse prior** on μ so that **posterior** for μ is approximately

$$(\mu|y) = (\mu|\bar{y}) \dot{\sim} N\left(\bar{y}, \frac{\sigma^2}{n}\right), \tag{10}$$

where $y = (y_1, \ldots, y_n)$.

Then **predictive distribution** for next observation is approximately

$$(y_{n+1}|y) = (y_{n+1}|\bar{y}) \dot{\sim} N\left[\bar{y}, \sigma^2\left(1 + \frac{1}{n}\right)\right], \tag{11}$$

and **log score**, ignoring linear scaling constants, is

$$LS(M_0|y) = \sum_{j=1}^{n} \ln p(y_j|y_{-j}), \tag{12}$$

where as before y_{-j} is y with observation j **set aside**.

But by **same reasoning**

$$p(y_j|y_{-j}) \doteq N\left(\bar{y}_{-j}, \sigma_n^2\right), \tag{13}$$

where $\bar{y}_{-j}$ is **sample mean** with observation j **omitted**, $\sigma_n^2 = \sigma^2\left(1 + \frac{1}{n-1}\right)$, so that

Log Score Approximation (continued)

$$\begin{aligned} \ln p(y_j|y_{-j}) &\doteq c - \frac{1}{2\sigma_n^2}(y_j - \bar{y}_{-j})^2 \quad \text{and} \\ LS(M_0|y) &\doteq c_1 - c_2 \sum_{j=1}^{n} (y_j - \bar{y}_{-j})^2 \end{aligned} \tag{14}$$

for some **constants** c_1 and c_2 with $c_2 > 0$.

Now it's **interesting fact** (related to behavior of **jackknife**), which you can prove by **induction**, that

$$\sum_{j=1}^{n} (y_j - \bar{y}_{-j})^2 = c \sum_{j=1}^{n} (y_j - \bar{y})^2 \tag{15}$$

for some $c > 0$, so finally for $c_2 > 0$ the **result** is that

$$LS(M_0|y) \doteq c_1 - c_2 \sum_{j=1}^{n} (y_j - \bar{y})^2, \tag{16}$$

i.e., in this model log score is **almost perfectly negatively correlated with sample variance**.

But in this model the **deviance** (minus twice the log likelihood) is

$$\begin{aligned} D(\mu) &= -2 \ln l(\mu|y) = c_0 - 2 \ln p(y|\mu) \\ &= c_0 + c_3 \sum_{j=1}^{n} (y_j - \mu)^2 \end{aligned} \tag{17}$$

for some $c_3 > 0$, encouraging suspicion that **log score should be strongly related to deviance**.

Deviance Information Criterion (DIC)

Given parametric model $p(y|\theta)$, Spiegelhalter et al. (2002) define **deviance information criterion** (DIC) (by analogy with other information criteria) to be estimate $D(\bar{\theta})$ of model (lack of) **fit** (as measured by deviance) plus **penalty for complexity** equal to twice **effective number of parameters** p_D of model:

$$DIC(M|y) = D(\bar{\theta}) + 2\,\widehat{p}_D, \tag{18}$$

where $\bar{\theta}$ is posterior mean of θ; they suggest that models with **low DIC** value are to be **preferred** over those with higher value.

When p_D is **difficult to read directly from model** (e.g., in **complex hierarchical models**, especially those with **random effects**), they motivate the following **estimate**, which is easy to compute from standard MCMC output:

$$\widehat{p}_D = \overline{D(\theta)} - D(\bar{\theta}), \tag{19}$$

i.e., difference between **posterior mean of deviance** and **deviance evaluated at posterior mean** of parameters (`WinBUGS` release 1.4 will **estimate** these quantities).

In **model** M_0, p_D is of course 1, and $\bar{\theta} = \bar{y}$, so

$$DIC(M_0|y) = c_0 + c_3 \sum_{j=1}^{n} (y_j - \bar{y})^2 + 2 \tag{20}$$

and conclusion is that

$$-DIC(M_0|y) \doteq c_1 + c_2 LS(M_0|y) \tag{21}$$

for $c_2 > 0$, i.e., (if this generalizes) **choosing model by maximizing log score and by minimizing DIC are approximately equivalent behaviors**.

(This connection was **hinted at** in discussion of Spiegelhalter et al. 2002 but never really made **explicit**.)

LS $\leftrightarrow$ DIC?

We're now (work in progress) **exploring the scope** of (21); in several **simple models** M so far we find for $c_2 > 0$ that

$$-DIC(M|y) \doteq c_1 + c_2 LS(M|y), \tag{22}$$

i.e., across repeated data sets generated from given model, even with small n DIC and LS can be **fairly strongly negatively correlated**.

Above argument **generalizes** to **any situation** in which **predictive distribution** is **approximately Gaussian** (e.g., **Poisson**(λ) likelihood with **large** λ, **Beta**(α, β) likelihood with **large** $(\alpha + \beta)$, etc.).

Example 3 continued. With **one-sample count data** (like number of hospitalizations in the T and C portions of IHGA data), people often choose between **fixed-** and **random-effects Poisson model formulations**: for $i = 1, \ldots, n$, and, e.g., with **diffuse priors**,

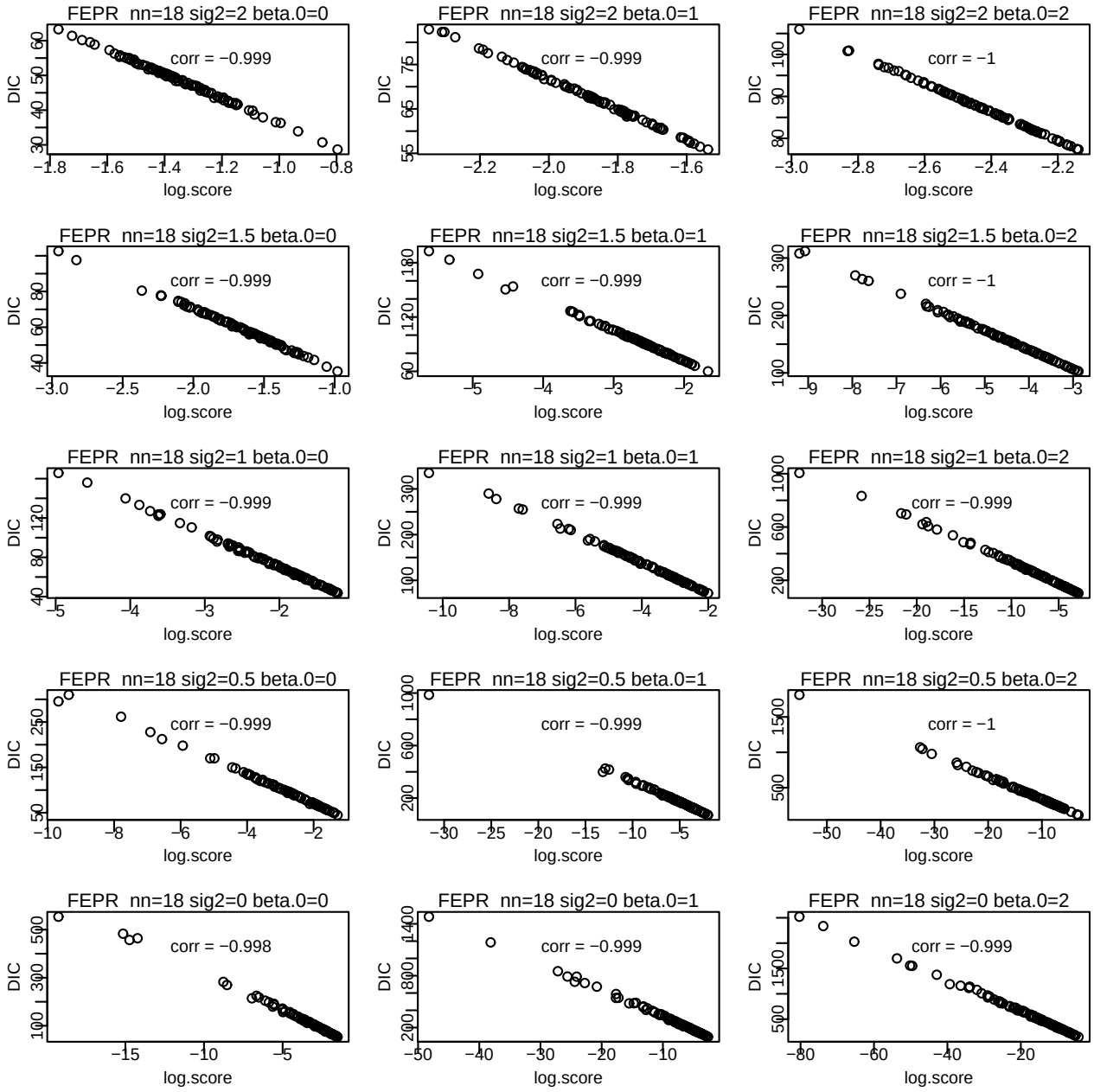
$$M_1\text{:} \left\{ \begin{array}{ccc} \lambda & \sim & p(\lambda) \\ (y_i|\lambda) & \overset{\text{IID}}{\sim} & \text{Poisson}(\lambda) \end{array} \right\} \quad \text{versus} \tag{23}$$

$$M_2\text{:} \left\{ \begin{array}{ccc} (\beta_0, \sigma^2) & \sim & p(\beta_0, \sigma^2) \\ (y_i|\lambda_i) & \overset{\text{indep}}{\sim} & \text{Poisson}(\lambda_i) \\ \log(\lambda_i) & = & \beta_0 + e_i \\ e_i & \overset{\text{IID}}{\sim} & N(0, \sigma^2) \end{array} \right\} \tag{24}$$

M_1 is **special case** of M_2 with $\left(\sigma^2 = 0, \lambda = e^{\beta_0}\right)$; likelihood in M_2 is **Lognormal mixture of Poissons** (often similar to fitting **negative binomial** distribution, which is **Gamma mixture of Poissons**).

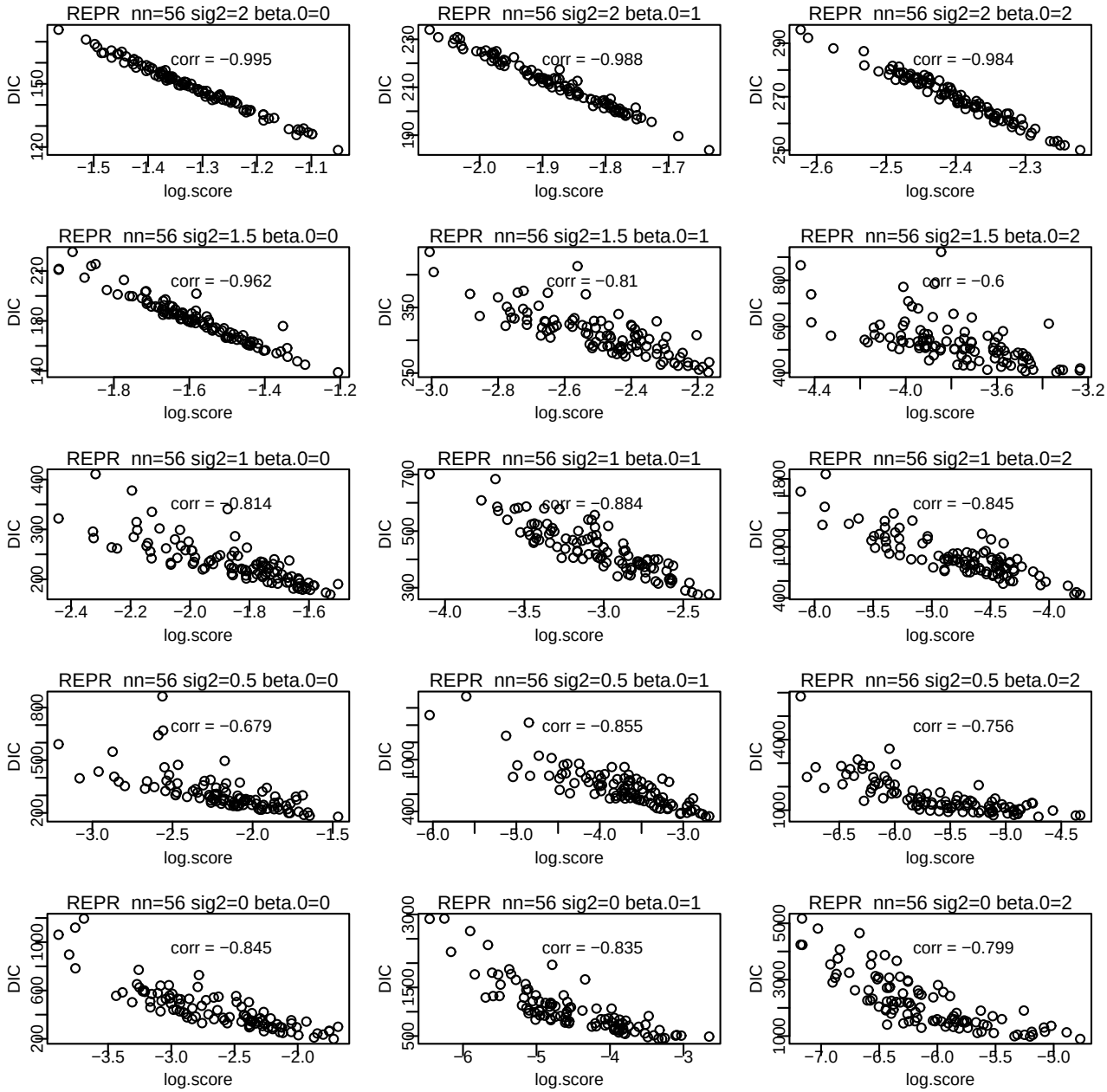
LS $\leftrightarrow$ DIC? (continued)

We conducted **partial-factorial simulation study** with factors $\{n = 18, 32, 42, 56, 100\}$, $\{\beta_0 = 0.0, 1.0, 2.0\}$, $\{\sigma^2 = 0.0, 0.5, 1.0, 1.5, 2.0\}$ in which (**data-generating mechanism**, **assumed model**) = $\{(M_1, M_1), (M_1, M_2), (M_2, M_1), (M_2, M_2)\}$; in each cell of this grid we used 100 **simulation replications**.



When **assumed model** is M_1 (**fixed-effects** Poisson), LS and DIC are **almost perfectly negatively correlated** (we have **mathematical explanation** of this).

LS $\leftrightarrow$ DIC? (continued)



When **assumed model** is M_2 (**random-effects** Poisson), LS and DIC are **less strongly negatively correlated** (DIC, at least as implemented in `WinBUGS` release 1.4, can **misbehave** with **mixture models**; see below), but **correlation increases with** n.

Example 3

As example of **correspondence between LS and DIC** in real problem, IHGA data were as follows:

Distribution of number of hospitalizations in IHGA study over two-year period:

Group	Number of Hospitalizations 0	1	2	3	4	5	6	7	n	Mean	SD
Control	138	77	46	12	8	4	0	2	287	0.944	1.24
Treatment	147	83	37	13	3	1	1	0	285	0.768	1.01

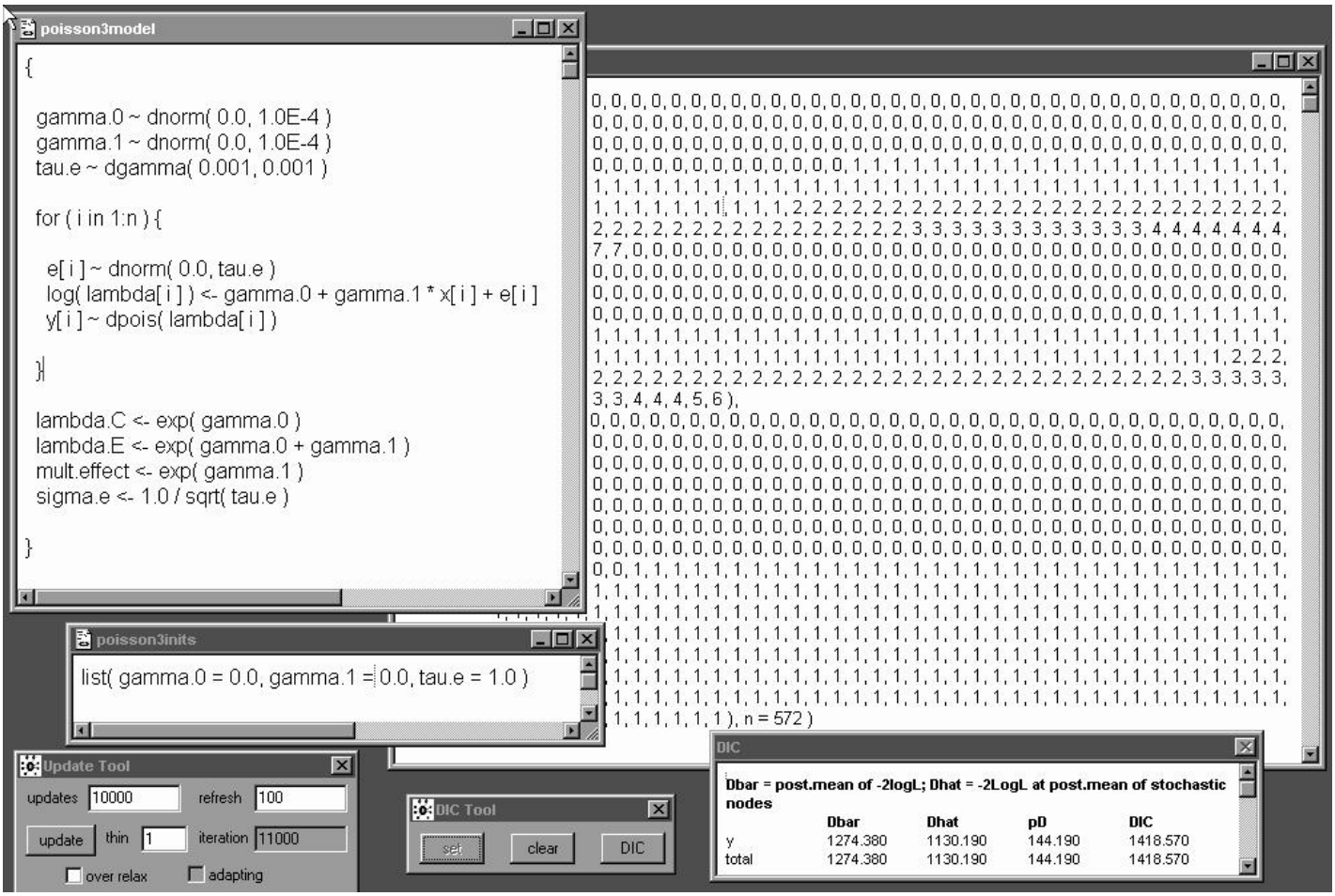
Evidently IHGA **lowered mean hospitalization rate** (for these elderly Danish people, at least) by $(0.944 - 0.768) =$ **0.176**, which is about $100\left(\frac{0.768-0.944}{0.944}\right) =$ **19%** reduction from control level, a difference that's **large in clinical terms**.

Four **possible models** for these data (not all of them good):

- **Two-independent-sample Gaussian** (diffuse priors);
- **One-sample Poisson** (diffuse prior), pretending treatment and control λs are equal;
- **Two-independent-sample Poisson** (diffuse priors), which is equivalent to **fixed-effects Poisson regression** (FEPR); and
- **Random-effects Poisson regression** (REPR), because C and T **variance-to-mean ratios** (VTMRs) are 1.63 and 1.32, respectively:

$$\begin{aligned} (y_i \,|\lambda_i) &\overset{\text{indep}}{\sim} \text{Poisson}(\lambda_i) \\ \log(\lambda_i) &= \beta_0 + \beta_1 x_i + e_i \\ e_i &\overset{\text{IID}}{\sim} N\left(0, \sigma_e^2\right) \\ \left(\beta_0, \beta_1, \sigma_e^2\right) &\sim \text{diffuse .} \end{aligned} \tag{25}$$

DIC Example



To use the **DIC feature** in `WinBUGS` to produce the screen shot above, I **fit** the REPR model as usual, did a **burn-in** of 1,000, **selected** `DIC` as a pull-down option from the `Inference` menu, **clicked** the `set` button in the `DIC Tool` window that popped up, **changed** the 1,000 to 10,000 in the `updates` window of the `Update Tool`, **clicked** `update`, and then **clicked** `DIC` in the `DIC Tool` when the monitoring run of 10,000 was finished—the `DIC` **results window** appears, with the `Dbar` ($\overline{D(\theta)}$), `Dhat` ($D(\bar{\theta})$), `pD` ($\hat{p}_D$), and `DIC` ($DIC(y)$) values.

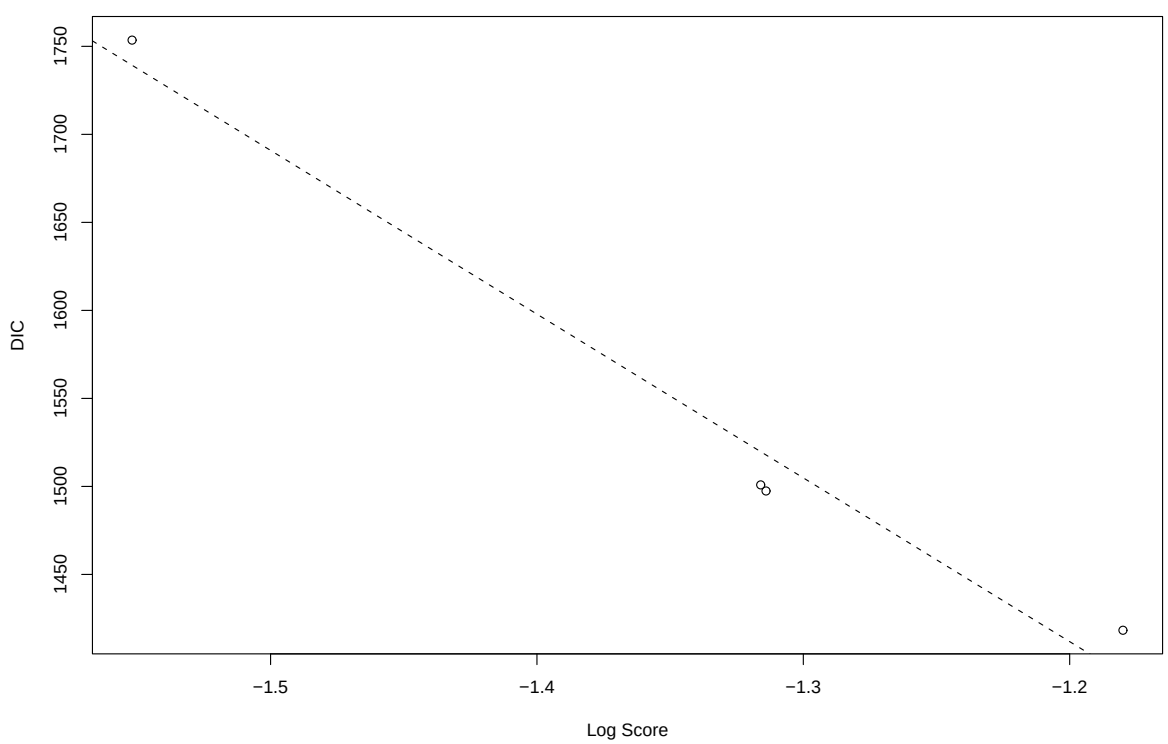
DIC Example (continued)

DIC and **LS** results on these four models:

Model	$\overline{D(\theta)}$	$D(\bar{\theta})$	$\widehat{p}_D$	$DIC(y)$	$LS(y)$
1 (Gaussian)	1749.6	1745.6	3.99	1753.5	−1.552
2 (Poisson, common λ)	1499.9	1498.8	1.02	1500.9	−1.316
3 (FEPR, different λs)	1495.4	1493.4	1.98	1497.4	−1.314
4 (REPR)	1275.7 1274.7 1274.4	1132.0 1131.3 1130.2	143.2 143.5 144.2	1418.3 1418.2 1418.6	−1.180

(3 REPR rows were based on **different monitoring runs**, all of length 10,000, to give idea of Monte Carlo noise level.)

As $\sigma_e \to 0$ in **REPR** model, you get **FEPR** model, with $p_D = 2$ parameters; as $\sigma_e \to \infty$, in effect **all subjects in study have their own** λ and p_D would be 572; in between at $\sigma_e \doteq 0.675$ (posterior mean), `WinBUGS` estimates that there are about **143 effective parameters in REPR model**, but its deviance $D(\bar{\theta})$ is so much lower that it **wins DIC contest** hands down.

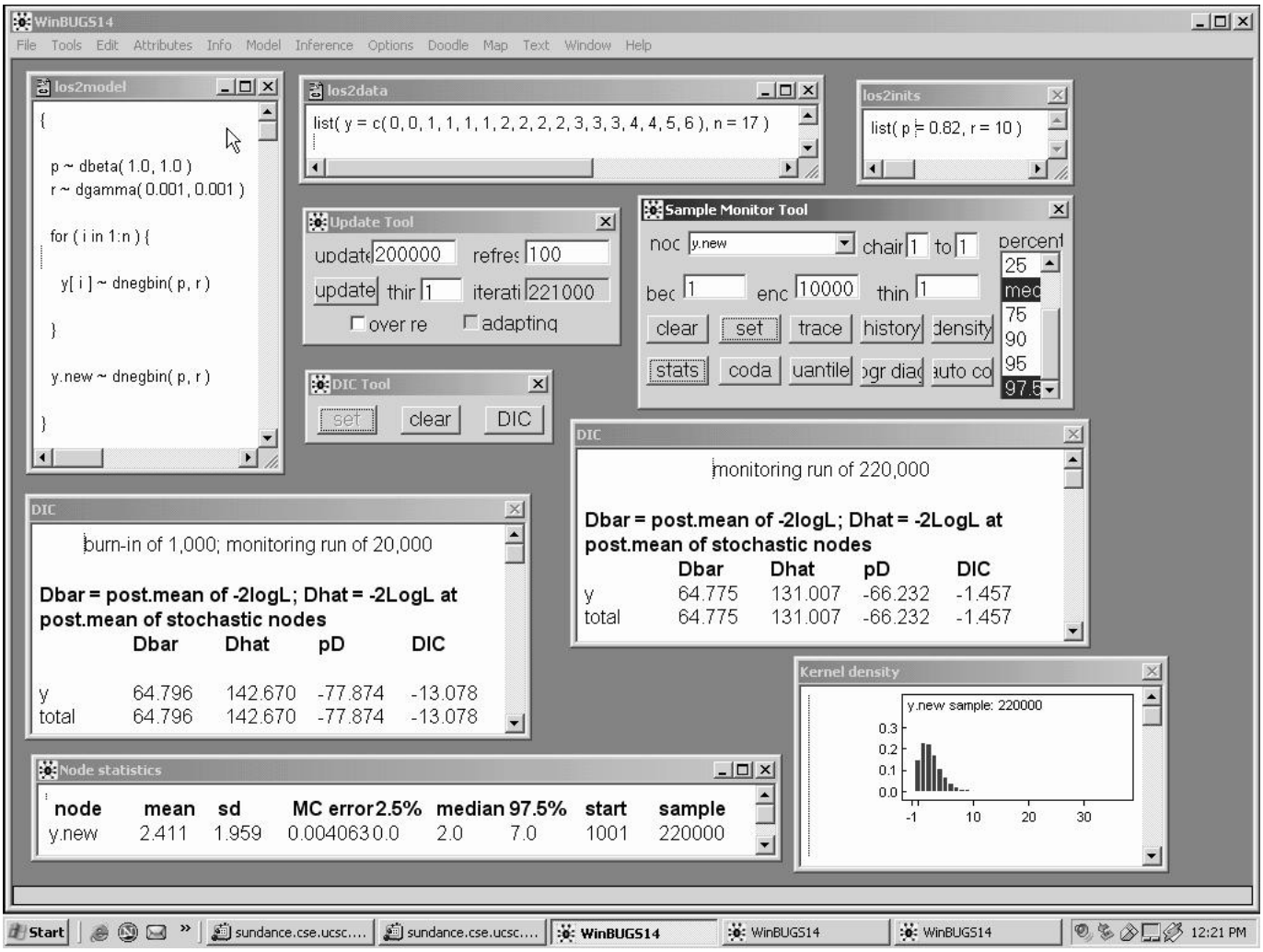


Correlation between LS and DIC across these four models is **−0.98**.

But WinBUGS DIC Can Misbehave

$y = (0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6)$ is a data set generated from the **negative binomial** distribution with parameters $(p, r) = (0.82, 10.8)$ (in WinBUGS notation); y has mean 2.35 and VTMR 1.22.

Using **standard diffuse priors** for p and r as in the BUGS examples manuals, the **effective number of parameters** p_D for the negative binomial model (which **fits the data quite well**) is estimated at **−66.2**, leading to a **DIC value** of **−1.5**:



The **online help** for WinBUGS release 1.4 warns about DIC misbehaving in **several contexts**, and WinBUGS release 1.4 **refuses to calculate DIC** in some models, but this seems to be a situation with **undocumented difficulties**.

7.4 LS Model Discrimination

On-going work: **DIC** calculated with **single MCMC run**; in one-sample setting with n observations and no closed form for predictive distribution, **brute-force LS** requires n **parallel MCMC runs**; how exactly is DIC approximating LS with **far less computation**?

Tentative conclusions: With **large** n (when LS can be **expensive**), **DIC** may provide **acceptable approximation** to model choice based on **LS**, but for **small** and **moderate** n **direct calculation** of **LS** may be **safer**.

With **LS utility function** and two models M_1 and M_2 to choose between, **MEU** says compute

$$\widehat{E}\left[U(M_i|y)\right] = \frac{1}{n}\sum_{j=1}^{n} \log p(y_j|M_i, y_{-j}) \tag{26}$$

and **choose** M_1 if $\widehat{E}\left[U(M_1|y)\right] \geq \widehat{E}\left[U(M_2|y)\right]$, otherwise M_2.

How **accurately** does this behavioral rule **discriminate** between M_1 and M_2?

Example: Recall that in **earlier simulation study**, for $i = 1, \ldots, n$, and with **diffuse priors**, we considered

$$M_1\text{:} \left\{ \begin{array}{ccc} \lambda & \sim & p(\lambda) \\ (y_i|\lambda) & \overset{\text{IID}}{\sim} & \text{Poisson}(\lambda) \end{array} \right\} \quad \text{versus}$$

$$M_2\text{:} \left\{ \begin{array}{ccc} (\beta_0, \sigma^2) & \sim & p(\beta_0, \sigma^2) \\ (y_i|\lambda_i) & \overset{\text{indep}}{\sim} & \text{Poisson}(\lambda_i) \\ \log(\lambda_i) & = & \beta_0 + e_i \\ e_i & \overset{\text{IID}}{\sim} & N(0, \sigma^2) \end{array} \right\}$$

LS Model Discrimination (continued)

As **extension** of previous simulation study, we generated data from M_2 and computed LS for models M_1 and M_2 in **full-factorial grid** $\{n = 32, 42, 56, 100\}$, $\{\beta_0 = 0.0, 1.0\}$, $\sigma^2 = 0.1, 0.25, 0.5, 1.0, 1.5, 2.0\}$, with **100** simulation replications in each cell, and monitored **percentages of correct model choice**.

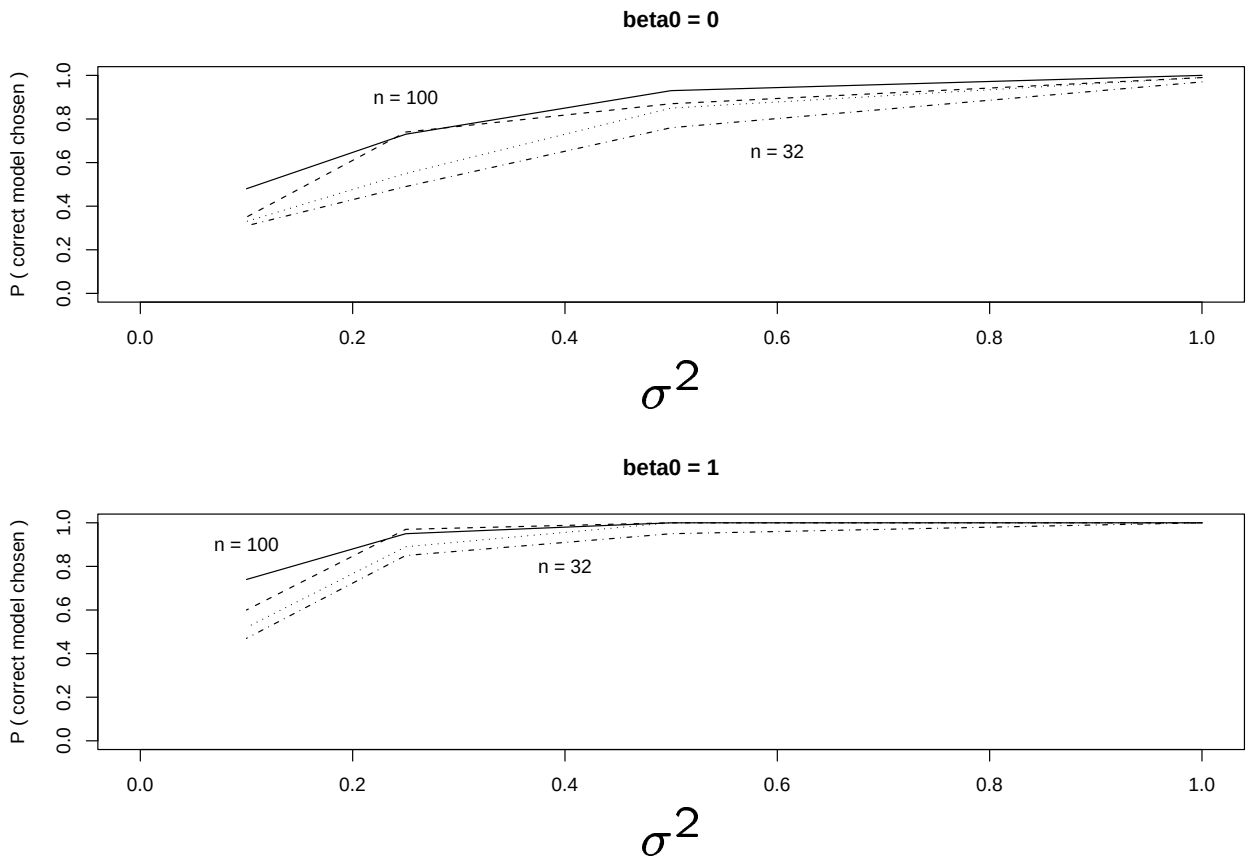
Examples of **results**:

n	σ^2	β_0	% Correct Decision	Mean Absolute Difference in LS
32	0.10	0	31	0.0009
32	0.10	1	47	0.0017
32	0.25	0	49	0.0016
32	0.25	1	85	0.0132
32	0.50	0	76	0.0171
32	0.50	1	95	0.2205
32	1.00	0	97	0.2369
32	1.00	1	100	4.0650
32	1.50	0	98	1.4370
32	1.50	1	100	17.4012
32	2.00	0	100	12.7938
32	2.00	1	100	63.9286

Even with n only **32**, the right model choice is made **more than 90% of the time** when $\sigma^2 > 0.5$ for $\beta_0 = 1$ and when $\sigma^2 > 1.0$ for $\beta_0 = 0$.

LS Model Discrimination (continued)

Graphical summary of these simulations (Bayesian decision-theoretic **power curves**):



NB Food for thought: In previous table **mean absolute LS difference** between M_1 and M_2 can be **tiny** even when this approach to model choice is **performing well**.

E.g., with $(n, \sigma^2, \beta_0) = (32, 0.25, 1)$, M_2 is correctly identified **85%** of the time, but its typical LS edge over M_1 is only about **0.013**—is this difference **large** in either **practical** or **statistical** terms?

7.5 Why Not Bayes Factors?

Much has been written about use of **Bayes factors** for **model choice** (e.g., Jeffreys 1935, Kass and Raftery 1995; excellent recent book by O'Hagan and Forster 2004 devotes almost **40 pages** to this topic).

After all, why not use **probability scale** to choose between M_1 and M_2?

$$\begin{aligned} \left[\frac{p(M_1|y)}{p(M_2|y)}\right] &= \left[\frac{p(M_1)}{p(M_2)}\right] \cdot \left[\frac{p(y|M_1)}{p(y|M_2)}\right] \\ \begin{pmatrix} \text{posterior} \\ \text{odds} \end{pmatrix} &= \begin{pmatrix} \text{prior} \\ \text{odds} \end{pmatrix} \cdot \begin{pmatrix} \text{Bayes} \\ \text{factor} \end{pmatrix} \end{aligned} \tag{27}$$

In fact, Kass and Raftery (1995) **note** that

$$\begin{aligned} \log \left[\frac{p(y|M_1)}{p(y|M_2)}\right] &= \log p(y|M_1) - \log p(y|M_2) \\ &= LS^*(M_1|y) - LS^*(M_2|y), \end{aligned} \tag{28}$$

where

$$\begin{aligned} LS^*(M_i|y) &\equiv \log p(y|M_i) \\ &= \log \left[p(y_1|M_i)\, p(y_2|y_1, M_i) \cdots p(y_n|y_1, \ldots, y_{n-1}, M_i)\right] \\ &= \log p(y_1|M) + \sum_{j=2}^{n} \log p(y_j|y_1, \ldots, y_{j-1}, M_i). \end{aligned}$$

Thus **log Bayes factor** equals **difference** between models in **something that looks like previous log score**, i.e., isn't previous procedure equivalent to choosing M_2 whenever the Bayes factor in favor of M_2 **exceeds 1**?

Out-of-sample LS $\neq$ BF

No; crucially, LS* is defined via **sequential** prediction of y_2 from y_1, y_3 from (y_1, y_2), etc., whereas LS is based on **averaging over all possible out-of-sample predictions**:

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

This distinction **really matters**: as is well known, with **diffuse priors** Bayes factors are **hideously sensitive** to particular **form** in which diffuseness is **specified**, but this defect is **entirely absent** from LS (and from other properly-defined **utility-based model choice criteria**).

(Various **attempts** have been made to **fix** this defect of Bayes factors, e.g., {partial, intrinsic, fractional} Bayes factors, well calibrated priors, conventional priors, intrinsic priors, expected posterior priors, ... (e.g., Pericchi 2004); all of these methods appear to require an appeal to **ad-hockery** which is not required by LS.)

Example: Integer-valued data $y = (y_1, \ldots, y_n)$;

$M_1 =$ **Geometric**(θ_1) likelihood with **Beta**(α_1, β_1) prior on θ_1;

$M_2 =$ **Poisson**(θ_2) likelihood with **Gamma**(α_2, β_2) prior on θ_2.

Out-of-sample LS $\neq$ BF

Bayes factor in favor of M_1 over M_2 is

$$\frac{\Gamma(\alpha_1+\beta_1)\Gamma(n+\alpha_1)\Gamma(n\bar{y}+\beta_1)\Gamma(\alpha_2)(n+\beta_2)^{n\bar{y}+\alpha_2}\left(\prod_{i=1}^n y_i!\right)}{\Gamma(\alpha_1)\Gamma(\beta_1)\Gamma(n+n\bar{y}+\alpha_1+\beta_1)\Gamma(n\bar{y}+\alpha_2)\beta_2^{\alpha_2}}.$$

Diffuse priors: take $(\alpha_1,\beta_1)=(1,1)$ and $(\alpha_2,\beta_2)=(\epsilon,\epsilon)$ for some $\epsilon>0$.

Bayes factor reduces to

$$\frac{\Gamma(n+1)\Gamma(n\bar{y}+1)\Gamma(\epsilon)(n+\epsilon)^{n\bar{y}+\epsilon}\left(\prod_{i=1}^n y_i!\right)}{\Gamma(n+n\bar{y}+2)\Gamma(n\bar{y}+\epsilon)\epsilon^{\epsilon}}.$$

This goes to $+\infty$ as $\epsilon\downarrow 0$, i.e., you can make the evidence in **favor** of the **Geometric model** over the **Poisson** as **large** as you want as a function of a quantity near 0 that **scientifically** you have **no basis** to specify.

By **contrast**

$$\begin{aligned} LS(M_1|y) &= \log\Big[\frac{(\alpha_1+n-1)\Gamma(\beta_1+s)}{\Gamma(\alpha_1+n+\beta_1+s)}\Big] \\ &\quad +\frac{1}{n}\sum_{i=1}^n \log\Big[\frac{\Gamma(\alpha_1+n-1+\beta_1+s_i)}{\Gamma(\beta_1+s_i)}\Big] \end{aligned}$$

and

$$\begin{aligned} LS(M_2|y) &= \frac{1}{n}\sum_{i=1}^n \log\Bigg[\frac{\Gamma(\alpha_2+s)}{\Gamma(y_i+1)\Gamma(\alpha_2+s_i)} \\ &\quad \cdot\Big(\frac{\beta_2+n}{\beta_2+n+1}\Big)^{\alpha_2+s_i}\Big(\frac{1}{\beta_2+n+1}\Big)^{y_i}\Bigg] \end{aligned}$$

and both of these quantities are **entirely stable** as a function of (α_1,β_1) and (α_2,β_2) near zero.

7.6 When Is A Model Good Enough?

LS method described here (not LS* method) can **stably** and **reliably** help in choosing between M_1 and M_2; but suppose M_2 has a (substantially) **higher LS** than M_1.

This doesn't say that M_2 is **adequate**—it just says that M_2 **is better than** M_1.

As mentioned above, a **full judgment of adequacy** requires **real-world input** (to what purpose will the model be put?), but you can answer a somewhat related question—**could the data have arisen from a given model**?—by **simulating** from that model many times, **developing** a distribution of (e.g.) LS values, and **seeing how unusual** the actual data set's log score is in this distribution (Draper and Krnjajić 2004).

However, this cannot be done **naively**, or result will be **poor calibration**.

Using **modification** of idea in Robins et al. (2000), we have developed method for **accurately calibrating log score scale**.

Inputs to our procedure: (1) A **data set** (e.g., with regression structure); (2) A **model** (can be parametric, non-parametric, or semi-parametric).

Simple example: data set $y = (1, 2, 2, 3, 3, 3, 4, 6, 7, 11)$, $n = 10$.

Given **model** $(*)$

$$\begin{aligned} (\lambda) &\sim \text{Gamma}(0.001, 0.001) \\ (y_i|\lambda) &\overset{\text{IID}}{\sim} \text{Poisson}(\lambda) \end{aligned} \tag{29}$$

Calibrating LS Scale

Step 1:

Calculate log score for this data set; say get LS $= -1.1$; call this **actual log score** (ALS).

Obtain posterior for λ given y based on this data set; call this **actual posterior**.

Step 2:

```
for ( i in 1:m1 ) {

  make a lambda draw from the actual posterior;
    call it lambda[ i ]

  generate a data set of size n from the second
    line of model (*) above, using
    lambda = lambda[ i ]

  compute the log score for this generated
    data set; call it LS[ i ]

}
```

Output of this loop is a vector of log scores; call this **V.LS**.

Locate ALS in distribution of LS values by computing percentage of LS values in V.LS that are $\leq$ ALS; call this percentage **unadjusted actual tail area** (say this is 0.22).

We know from our own simulations and the literature (e.g., Robins et al. 2000) that this tail area is **conservative**, i.e., with the 0.22 example above an adjusted version of it that is well calibrated would be **smaller**.

Calibrating LS Scale (continued)

Gelman, Meng and Stern (1996) proposed this unadjusted version without specifying a particular **discrepancy function** (the predictive log score in our case), but Robins et al. (a) demonstrated that Gelman et al. approach was **poorly calibrated** and (b) suggested some ways to produce improved calibration.

We've **modified** and implemented one of the ways suggested by Robins et al., and we've shown that it does indeed work even in rather small-sample situations, although our approach to implementing the basic idea is quite **computationally intensive**.

Step 3:

```
for ( j in 1:m2 ){

  make a lambda draw from the actual posterior;
    call it lambda*.

  generate a data set of size n from the second line
    of model (*) above, using lambda = lambda*;
    call this the simulated data set

  repeat steps 1, 2 above on this
    simulated data set

}
```

The result will be a vector of unadjusted tail areas; call this **V.P**.

Compute the percentage of tail areas in V.P that are $\leq$ the unadjusted actual tail area; this is the **adjusted actual tail area**.

Calibrating LS Scale (continued)

The claim is that the 3-step procedure above is **well-calibrated**, i.e., if the sampling part of model (*) really did generate the observed data, the distribution of adjusted actual tail areas obtained in this way would be **uniform**, apart from simulation noise.

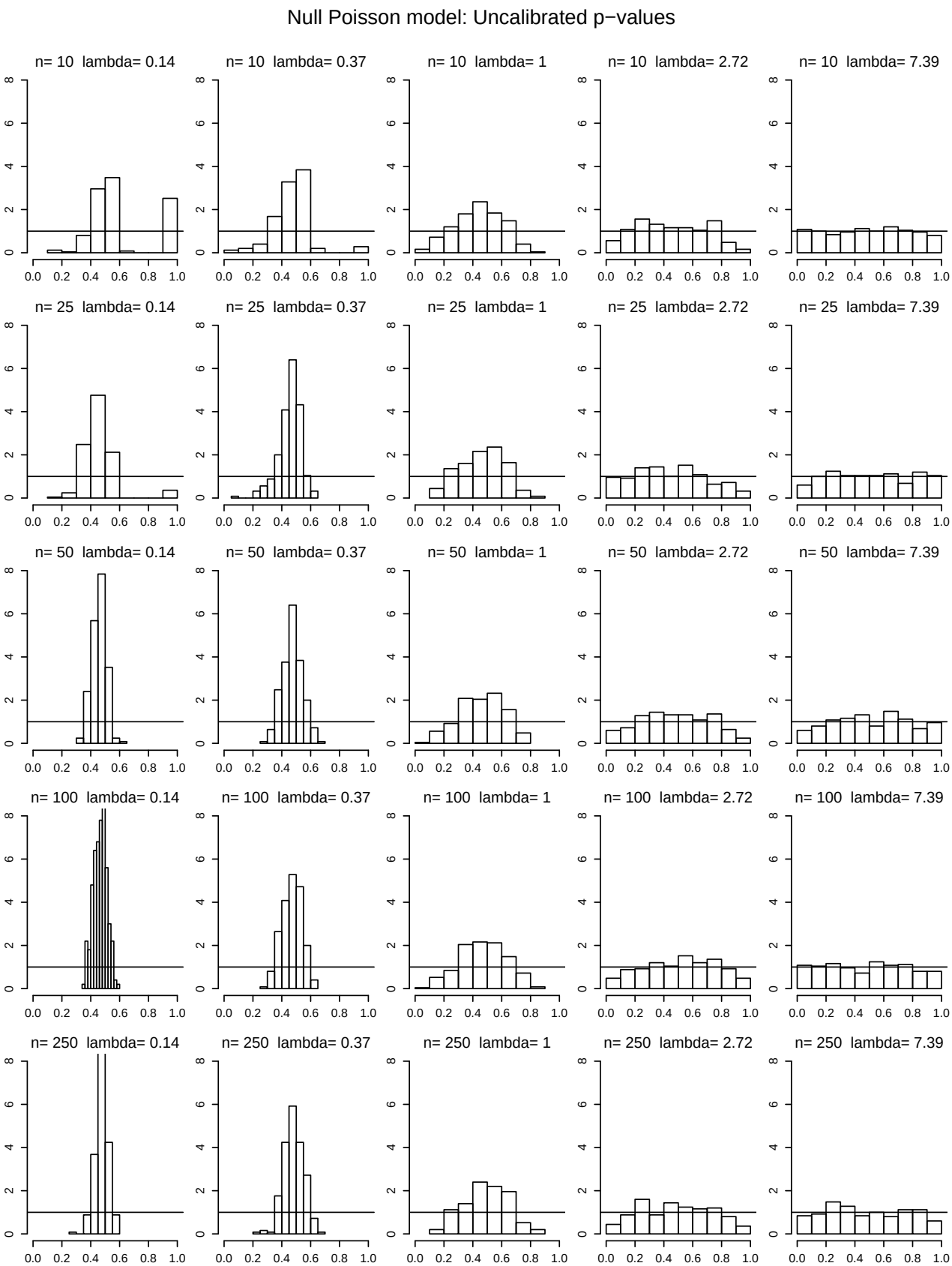
This claim can be tested by building a **big loop** around steps 1–3 as follows:

```
Choose a lambda value of interest; call it lambda.sim

for ( k in 1:m3 ) {

  generate a data set of size n from the
    second line of model (*) above, using
    lambda = lambda.sim; call this the
    validation data set

  repeat steps 1-3 on the validation data set

}
```

The result will be a vector of **adjusted P-values**; call this **V.Pa**.

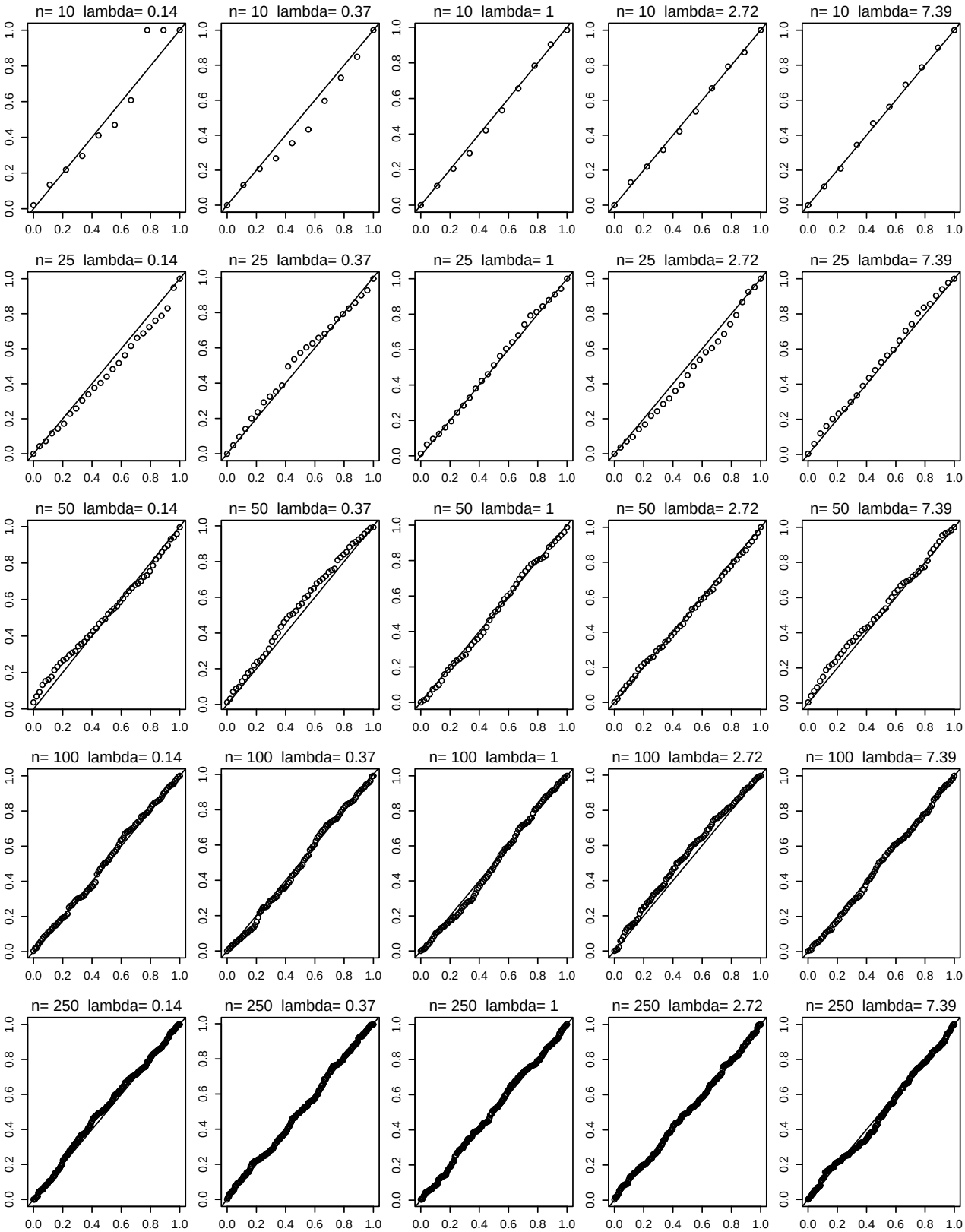
We have **verified** (via simulation) in several simple (and some less simple) situations that the values in V.Pa are close to $U(0, 1)$ in distribution.

Uncalibrated p-values

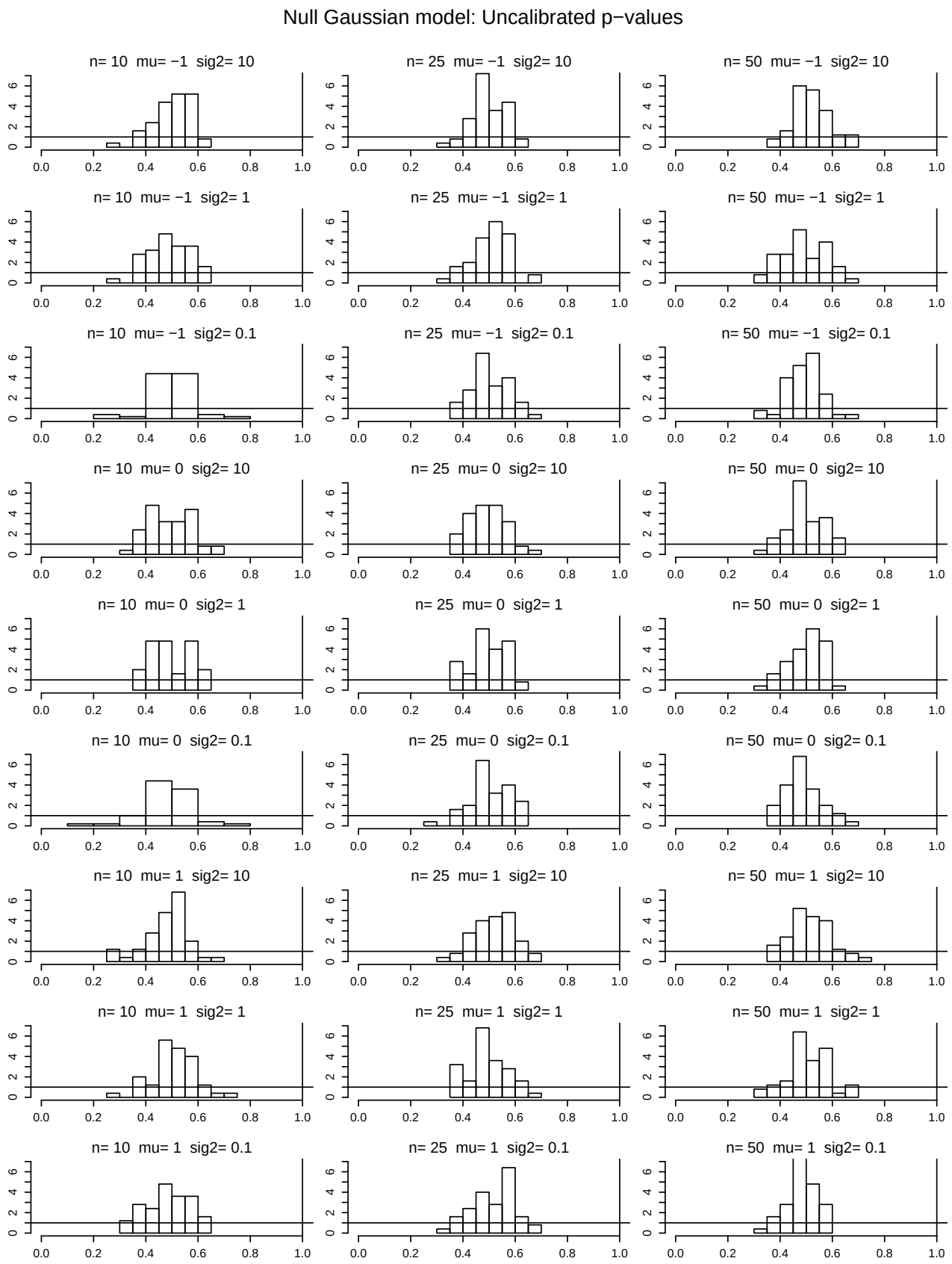


Calibrated p-values

Null Poisson model: Calibrated p−values vs uniform(0,1)

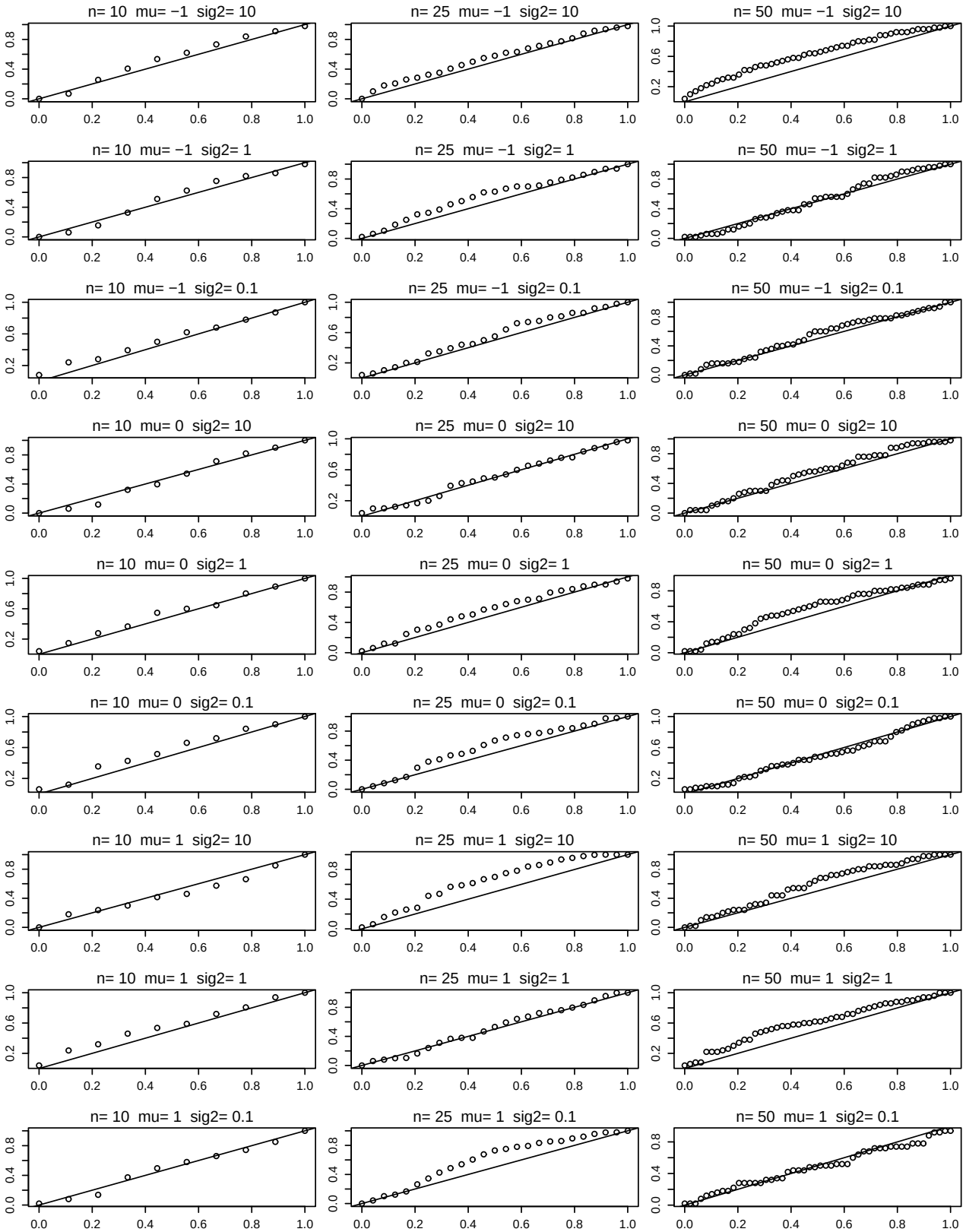


Uncalibrated p-values



Calibrated p-values

Null Gaussian model: Calibrated p−values vs uniform(0,1)



7.7 Conclusions

• {**Exchangeability judgments** plus **non-parametric** and **semi-parametric** modeling} = **Bayesian model specification** in many problems, but we need to develop a **lot more experience** in **practical problems** in putting **distributions** on **functions**.

• **Three-way cross-validation** (3CV) is another way to avoid the **dilemma** posed by **Cromwell's Rule** in Bayesian model specification.

• Model choice is really a **decision problem** and should be approached via **MEU**, with a utility structure that's **sensitive to the real-world context**.

• **Leave-one-out predictive log score** (LS) has a sound utility basis, and can yield **stable** and **accurate** model specification decisions.

• DIC is a **good** and **fast** approximation to LS in **some models**, but sometimes it **misbehaves**.

• **Bayes factors** are a **bad choice** for model specification when context suggests **diffuse** priors.

• The basic Gelman et al. (1996) method of posterior predictive model-checking is **poorly calibrated**: when it gives you a tail area of, e.g., **0.22**, the calibrated equivalent may well be **0.02**.

• We have modified an **approach** suggested by Robins et al. (2000) to help answer the question "Could the data have arisen from model M?" in a **well-calibrated** way.

Odds and Ends and Plugs

1 I've created a **web page** for this short course called `www.ams.ucsc.edu/~draper/Boston2004.html` with the **handouts** in downloadable form; **data sets** and `WinBUGS` **code** from the case studies are also available there.

2 Bill Browne and I are planning to give a **3–day course** on **multilevel modeling** (especially via **MCMC**) using `MLwiN`, probably in **California**, probably sometime in **2005**.

3 My UCSC colleague **Thanasis Kottas** (an expert on Bayesian nonparametric modeling) and I will co-present the **30th annual Summer Institute of Applied Statistics** at Brigham Young University in June 2005 on **practical Bayesian nonparametrics**: this will be a **2.5–day course** with **case studies** and **hands-on computing experience**.

2–3 will be **announced** on (e.g.) `bayes-news`, posted to the `BUGS` **email user group**, etc.

References

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

Dey D, Mueller P, Sinha D (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer Verlag (Lecture Notes in Statistics, Volume 133).

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

de Finetti B (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré*, **7**, 1–68.

de Finetti B (1938). Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles*, **739**.

de Finetti B (1970). *Teoria della Probabilità*, Volumes 1 and 2. Turin: Einaudi.

Draper D (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B*, **57**, 45–97.

Draper D (1996). Utility, sensitivity analysis, and cross-validation in Bayesian model-checking. *Statistica Sinica*, **6**, 760–767 (discussion of "Posterior predictive assessment of model fitness via realized discrepancies," by A Gelman, X-L Meng, and H Stern).

Draper D, Fouskakis D (2000). A case study of stochastic optimization in health policy: problem formulation and preliminary results. *Journal of Global Optimization*, **18**, 399–416.

Draper D, Fouskakis D (2004). Stochastic optimization methods for cost-effective quality assessment in health. Submitted.

Draper D, Krnjajić M (2005). Three-way cross-validation for well-calibrated model exploration. In preparation.

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

Fouskakis D, Draper D (2002). Stochastic optimization: a review. *International Statistical Review*, **70**, 315–349.

References (continued)

Geisser S, Eddy WF (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, **74**, 153–160.

Gelfand AE, Dey DK, Chang H (1992). Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors), Oxford: Oxford University Press, 147–167.

Gelman A, Meng X-L, Stern H (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistica Sinica*, **6**, 733–760.

Good IJ (1950). *Probability and the Weighing of Evidence*. London: Griffin.

Hendriksen C, Lund E, Stromgard E (1984). Consequences of assessment and intervention among elderly people: a three year randomised controlled trial. *British Medical Journal*, **289**, 1522–1524.

Jeffreys H (1939). *Theory of Probability*. Oxford: Oxford University Press.

Kass RE, Raftery AE (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Key JT, Pericchi LR, Smith AFM (1998). Bayesian model choice: what and why? (with discussion). In *Bayesian Statistics 6*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (editors). Oxford University Press, 343–370.

O'Hagan A, Forster J (2004). *Bayesian Inference*, second edition. London: Arnold.

Pericchi L (2004). Model selection and hypothesis testing based on objective probabilities and Bayes factors. Manuscript.

Robins JM, van der Vaart A, Ventura V (2000). Asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association*, **95**, 1143–1156.

Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B*, **64**, 583–616.

Walker S, Damien P, Lenk P (2004). On priors with a Kullback-Leibler property. *Journal of the American Statistical Association*, **99**, 404–408.