Bayesian Sample Size Computations in Complex Models with Application to Repeated Measures Random Effects Model Design*

Robert Weiss and Yan Wang
Department of Biostatistics
UCLA School of Public Health
Los Angeles CA 90095-1772 U.S.A.

September 3, 1997

Abstract

Bayesian methodology is developed to chose the sample size in complex problems where testing a null hypothesis is of interest. The approach permits propagation of uncertainty in quantities which are unknown, and permits computation of power and type I error. A graphical diagnostic is used to assess the sensitivity of the design to model specification and sample size specification. The sample size is chosen large enough to provide a pre-specified probability that the Bayes factor between the null and alternative hypothesis is larger than a cut-off. We develop methodology for models with covariates with uncertain distributions and treatments, to multivariate, unbalanced and missing response data.

We apply the methodology to a repeated measures random effects model with a predictive prior based on data from an earlier study.

Key Words: Bayes Factor; Experimental Design; Hierarchical Model; Prior Predictions.

*This work was supported by grant #GM50011 from the National Institute for General Medical Science of the NIH.

1 Introduction

The goal of a medical study is often stated as an interest in testing a null versus an alternative hypothesis. Previously, Bayesian methodology has been used to choose the sample size to guarantee frequentist power for a classical test of hypothesis (Freedman and Spiegelhalter 1983; Spiegelhalter and Freedman 1986) and to choose the sample size to reduce the posterior variance to a prespecified size or to guarantee a fixed level of posterior coverage in an interval of specific size. For references and a review of Bayesian sample size selection methods, see Adcock (1997). Müller and Parmigiani (1995) is quite close in spirit to the current work. They give examples of sample size specification for estimating a binomial probability and for a survival analysis where parameter estimation is of interest. They minimize various loss functions to choose the sample size and use Monte Carlo simulations to estimate the utilities of various sample sizes. However, they do not give methodology for testing a hypothesis.

The Bayesian tool for testing a hypothesis is a Bayes factor. Weiss (1997) and Verdinelli (1996) independently introduced the idea of choosing the sample size to guarantee that the Bayes factor is larger than a certain prespecified size. Both illustrated the methodology in the case of a simple null hypothesis $H_0 : \mu = 0$ versus $H_a : \mu \neq 0$ for independent and identically distributed data $y_i|\mu \sim N(\mu, \tau)$, for τ known.

Algebraically, let Y denote the data, let H_0 and H_1 be the competing hypotheses with parameters θ_k, sampling distributions $f(Y|\theta_k, H_k)$ and priors $p(\theta_k|H_k)$ for $k = 0, 1$. Then $f(Y|H_k) = \int f(Y|\theta, H_k)p(\theta|H_k)d\theta_k$ for $k = 0, 1$ are the prior predictive distributions. The Bayes factor is

$$B_{01} = \frac{f(Y|H_0)}{f(Y|H_1)}$$

and define $b_{01} = \log B_{01}$ as the log Bayes factor in favor of H_0 against H_1. Kass and Raftery (1995) suggest b_{01} greater than $+3$ or less than -3 constitute *strong* support for or against H_0 respectively; and $b_{01} > 5$ or $b_{01} < -5$ is said to constitute *very strong* support for H_0 or H_1 respectively.

The goal of the design is to select the sample size n and possibly other aspects of the design so that the prior predictive probabilities $p_0(a_0) = P(b_{01} > a_0|H_0)$ and/or $p_1(-a_1) = P(b_{01} < -a_1|H_1)$ are suitably large for some $a_0, a_1 \geq 0$. We explore choices of $a_0 = a_1$ equal to 3 or 5 in our example. Instead of specifying cutoffs a_k on b_{01}, another approach is to choose n so that the posterior probability that H_k is greater than b_k is large for $k = 0, 1$ and $0 < b_k < 1$. Let π_k be the prior probability of H_k then taking $a_k = \log[(1-b_k)\pi_1/(b_k\pi_0)]$ accomplishes this and leads to the same methodology as before. Verdinelli (1996) chooses n to make $\pi_0 p_0(\log 3) + \pi_1 p_1(-\log 3)$ larger than a pre-specified probability. In the absence of prior probabilities $p(H_k)$, we might pick n to make the sum $p_0(3)+p_1(-3) \geq a^*$, with $0 < a^* < 2$. This effectively takes $p(H_k) = .5$.

An alternative method of sample size specification is to find $\psi_{.05}(n)$, the 5^{th} percentage point of the distribution $p(b_{01}|H_0)$, as a function of n, and choose n so that $P(b_{01} < \psi_{.05}(n)|H_1) > .8$ (Weiss 1997). This treatment of the Bayes factor is classical in nature. Specifically, it is an empirical Bayes approach; use a Bayesian prior at all levels below the top level to integrate out so-called nuisance parameters. Then use a classical approach at the top level. In the testing paradigm, the top level is the two hypotheses. The hypotheses are of interest, and parameter estimates are not of interest, although obviously a careful data analysis would report posterior estimates.

Unlike the situations studied in Weiss (1997) and Verdinelli (1996), it is often impossible to algebraically calculate the prior predictive distribution of b_{01} given H_0 or H_1. We then use a combination of Monte Carlo simulation, algebraic calculations and numerical integration to study $p(b_{01}|H_k)$, $k = 0, 1$ for a complex model and set of hypotheses. Specification of appropriate priors is the subject of considerable ongoing research; we use a predictive prior (Weiss, Wang and Ibrahim 1997) based on a previous experiment as the prior under H_k. In the next section we discuss some issues involved with the modeling and the calculation of Bayes factors. In section 3, we outline a general algorithm to simulate the distributions $p(b_{01}|H_k)$, $k = 0, 1$ for design problems with covariates, missing and multivariate data. Section 4 illustrates the problem of choosing a sample size for a complicated hierarchical repeated measures data random effects model based on a prior study. The paper closes with a short discussion.

2 Covariates, Missing Data and Bayes factors

We assume a proper prior is available for the parameters θ_k under the respective hypotheses H_k. However, this is not enough; in a general situation, the sampling density of Y depends not only on parameters θ but also on covariates X which may or may not be completely under the control of the experimenter and whose distribution may or may not be fully known. A model for the covariate density is needed to properly simulate the prior predictive distribution of the Bayes factors under H_k. Some covariates are simple to model as for example a randomized binary treatment indicator, which can be modeled as a Bernoulli random variable with probability of success equal to $\pi = .5$. Continuous covariates such as age or disease score can be more complicated to model, although a transformation, such as for example, an inverse empirical cumulative distribution function or a power transformation, may permit them to be modeled either as uniform or normal random variables. A kernel density estimate might be used to estimate a low dimensional density. Correlated covariates may in general be difficult to model.

In designing a study to be analyzed with a nontrivial statistical model, it is important to take uncertainty in the covariate values into account when assessing the uncertainty in outcomes of the proposed experiment. When we make an assumption that X has density $g(X)$ when the distribution is unknown conditions on information not actually held. Fixing a known set of covariates or distributions results in summaries about the possible outcomes of the proposed experiment that are conditional. Unconditional predictions from a particular sample size about possible experimental results are particularly important for funding agencies and the experimenter in assessing the potential value of a particular proposed sample size. Occasionally conditional predictions can also be of interest when interest lies in the properties of a design conditional on some important event, although that event is most likely to involve the parameters or predictions of interest and not the covariates or their distributions.

When a sample x_j, $j = 1, \ldots, J$ of continuous m-dimensional covariates is available as prior information, a simple approach to modeling the joint distribution is to use a kernal density estimator. Let $H(x)$ be the m-dimensional kernel; assume $H(x)$ is easy to draw from. Then a draw from the kernel density estimate is generated by first drawing x from $H(x)$, and selecting a point x_j from the set of previously observed X's at random. The draw from the kernel density is $x + x_j$. For choice of kernels, see Scott (1992).

In general, we model the sampling distribution $g(X|\phi)$ for the covariates given unknown parameters ϕ. We set up a prior $q(\phi)$ for ϕ using preliminary data and substantive knowledge about the population. Often, a previous experiment or other information source for the priors $p(\theta_k|H_k)$ will additionally provide information about $g(X|\phi)$ and $q(\phi)$. Assuming that ϕ and $X|\phi$ are independent of θ_k and k the index of the hypothesis, we have

Result 1.

$$
\begin{aligned}
B_{01} &= \frac{\int\int f(Y|X,\theta_0,H_0)g(X|\phi)q(\phi)p(\theta_0|H_0)d\phi d\theta_0}{\int\int f(Y|X,\theta_1,H_1)g(X|\phi)q(\phi)p(\theta_1|H_1)d\phi d\theta_1} \\
&= \frac{\int f(Y|X,\theta_0,H_0)p(\theta_0|H_0)d\theta_0 g(X)}{\int f(Y|X,\theta_1,H_1)p(\theta_1|H_1)d\theta_1 g(X)}, \qquad (1)
\end{aligned}
$$

and the Bayes factor does not involve the distribution of $X|\phi$ or the prior of ϕ, since $g(X) = \int g(X|\phi)p(\phi)d\phi$ factors out of both the numerator and denominator.

For multivariate response data where the n responses form an n by d matrix with rows conditionally independent, we would usually model the data Y and then the missingness indicator $R|Y$, with R also n by d. Provided the data is assumed missing at random (MAR, see Little and Rubin 1987, p. 14) and the distribution of the missingness does not depend on the model, then the Bayes factor does not depend on the value of R nor on the model for the missingness. In particular, split the responses Y into $(Y_{\text{obs}}, Y_{\text{mis}})$, the observed and missing portions of the data. The distribution of R can depend on X, Y_{obs}, and unknown parameters ψ independent of θ_k under H_k. Let $h(R|Y_{\text{obs}}, X, \psi)r(\psi)$ be the density and prior for R and ψ. Then we have the following result for the missingness indicator which parallels (1) for covariates.

Result 2.

$$
\begin{aligned}
B_{01} &= \frac{\int h(R|Y_{\text{obs}}, X, \psi) r(\psi) f(Y_{\text{mis}}|Y_{\text{obs}}, X, \theta_0, H_0) f(Y_{\text{obs}}|X, \theta_0, H_0) p(\theta_0|H_0) d\psi dY_{\text{mis}} d\theta_0}{\int h(R|Y_{\text{obs}}, X, \psi) r(\psi) f(Y_{\text{mis}}|Y_{\text{obs}}, X, \theta_1, H_1) f(Y_{\text{obs}}|X, \theta_1, H_1) p(\theta_1|H_1) d\psi dY_{\text{mis}} d\theta_1} \\
&= \frac{\int f(Y|X, \theta_0, H_0) p(\theta_0|H_0) d\theta_0 h(R|Y_{\text{obs}}, X)}{\int f(Y|X, \theta_1, H_1) p(\theta_1|H_1) d\theta_1 h(R|Y_{\text{obs}}, X)}. \qquad (2)
\end{aligned}
$$

In line 1 of equation (2), the distribution of Y_{mis} integrates out of both numerator and denominator and then the marginal $h(R|Y_{\text{obs}}, X)$ factors out of the numerator and denominator. If the distribution of R had depended on Y_{mis}, then the density $h(R|Y_{\text{obs}}, Y_{\text{mis}}, X)$ of R after integrating out ψ would depend on Y_{mis} as well as Y_{obs} and X. Then we would not be able to integrate in (2) in closed form with respect to Y_{mis}, nor would the cancelation of $h(R|Y_{\text{obs}}, X)$ occur. Thus we would need to integrate with respect to Y_{mis} in the numerator and denominator of (2) and the calculation of the Bayes factor would be even more difficult than it currently is. In the algorithm in the next section we let R depend only on ψ and X.

Calculations of Bayes factors often require difficult numerical integrations. For the repeated measures random effects model Weiss, Wang and Ibrahim (1997) give a procedure for calculating the Bayes factors between models with different sets of fixed effects. We use a modification of this procedure for our sample size specification in section 4; details are given in the appendix. General numerical procedures for calculating Bayes factors are the subject of much current research, for example, Chen and Shao (1997a, 1997b), Chen, Ibrahim and Yiannoutsos (1996), Chib (1995), Gelman and Meng (1994), Geyer (1994), Meng and Wong (1994), Newton and Raftery (1995), Weiss (1996).

3 Simulating the predictive distributions of the Bayes factor

After specifying the necessary distributions, simulating $p(b_{01}|H_k)$ for a given sample size involves the following steps. For each of $l \in \{1, \ldots, L_k\}$ times

1. Sample $\phi^{(l)}$, the parameters of the X distribution from $q(\phi)$.
2. Sample the covariates $X^{(l)}$ from $g(X|\phi^{(l)})$.
3. Sample $\psi^{(l)}$, the parameters of the missingness distribution from $r(\psi)$.
4. Sample missingness indicator variable $R^{(l)}$ from $R|\psi^{(l)}, X^{(l)}$.
5. Sample the unknown parameters $\theta_k^{(l)}$ from the prior $p(\theta|H_k)$.
6. Sample the data $Y_k^{(l)}$ from $p(Y|X^{(l)}, \theta_k^{(l)}, H_k)$.
7. Calculate the Bayes factor $b_{01,k}^{(l)} = \log[f(Y_{\text{obs},k}^{(l)}|H_0, R)/f(Y_{\text{obs},k}^{(l)}|H_1, R)]$ based on the sampled observed data.

For notational simplicity a subscript identifying the sample size n and other design parameters is omitted above. It may be possible to only sample Y_{obs} rather than the entire Y vector. It may also be possible to reduce the computations by reusing $\phi^{(l)}$, $X^{(l)}$, $\psi^{(l)}$, and $R^{(l)}$ for simulations under both H_0

and H_1. Having calculated samples from $p(b_{01}|H_k)$, we estimate the cutoff point and power (classical approach), or the sum $P(b_{01} > a_0|H_0) + P(b_{01} < -a_1|H_1)$, (Bayesian approach); or other summary statistic as desired, depending on the specific utility function used in designing the study.

To identify the needed sample size, we use a simple search, taking $n = 20, 30, 40, \ldots$ until the sample size is bracketed. More sophisticated search routines can be used especially when the Bayes factor calculations are time consuming. For early calculations, we take L_k small, such as 100, which gives a standard error of approximately $.05 = (.5^2/100)^{1/2}$ for the estimated probability of interest. We increase L_k after tentatively identifying the needed sample size. Kernel density estimates of $p(b_{01}|H_k)$ are checked at each step. We check if and how the $p(b_{01}|H_k)$ change with increasing sample size, and whether substantial probability runs off towards $\pm\infty$ with increasing sample size. We also look for the shape of the distributions, presence of long tails, skewness, and multiple modes.

4 Sample Size for a Repeated Measures Pediatric Pain Study

Classical sample size calculations for repeated measures analyses usually rely on possibly non-central F or χ^2 approximations to the distribution of classical test statistic under the null and alternative hypotheses (Lui and Cumberland

1992; Rochon 1991; Muller, LaVange, Ramey and Ramey 1992). Most research has assumed either 2 or sometimes more than two groups with equal sample sizes. Most calculations are restricted to a fixed X design, and methods to allow for random covariates are either unavailable or at best restricted to a few special cases. Similarly, methods that allow for missing data are unavailable (confer Muller et al 1992, section 3.2). Standard procedure to adjust for loss of cases seems to be to estimate the sample size assuming no missing data. Then given a point estimate $0 < \omega < 1$ of the expected fraction of missing data, allowance for missing data typically takes the form of inflating the desired sample size by $(1-\omega)^{-1}$. No adjustment seems available for missing observations within a case. For a general discussion of issues in sample size calculations for repeated measures data, see Muller et al (1992).

4.1 Context of the Design

In this example we consider the design of a followup study based on a prior repeated measures data set with missing data. Observations are the log of the time in seconds that a child can keep his or her hand in quite cold water before being forced to remove it. The time is a proxy for pain tolerance. The prior study had two covariates, Coping Style (CS) and Treatment (TMT). If treatment has an effect, then CS and TMT are thought to interact. The CS is observed and is not under the control of the investigator. The CS is either attend (A) or distract (D). Attenders pay attention to their arm in the cold water or the experimental apparatus, while distractors think about other things such as a vacation or schoolwork. The TMT was randomized as either counseling to attend (A), distract (D), or a null treatment (N). The prior study design had $m_1 = 3$ baseline observations followed by the counseling intervention, and then $m_2 = 1$ response observations.

The hypotheses of interest are H_0: no treatment effect against H_1: a treatment effect which may be different for attenders and distracters. Analysis by Weiss, Wang and Ibrahim (1997) of the original study using a predictive prior and the 58 complete data cases indicated that the data strongly supported H_0 against H_1. This was somewhat surprising, given that other analyses (Fanurik, Zeltzer, Roberts and Blount 1993; Weiss 1994) supported H_1. Thus it is of interest to design a followup study to discover using Bayesian methods whether the treatment intervention does indeed have an effect on pain tolerance. Since there is little interest in the null treatment, in the followup study we eliminate the N treatment.

The four trials of the original study were given on two days, approximately two weeks apart. No effect due to days has been found. In the followup, it is planned to have three trials per day instead of two with the intervention taking place before the fourth or fifth trial. A desire for balance suggests having the intervention before the fourth trial. Efficiency considerations might put the intervention before the third trial, but this is probably not practicable, since the effect of the two week break between trial 3 and 4 on the intervention efficacy is unknown and is unlikely to be ignorable. As long as the intervention takes place on day two, we are comfortable assuming that the mean structure is the same across trials before intervention, changes because of the intervention, and then remains constant again. The most reasonable design for the followup study will have $m_1 = 3$ pre-treatment trials and $m_2 = 3$ post-treatment trials; we investigate the effects of taking $m_1 = 4$ and $m_2 = 2$ and $m_1 = 2$ and $m_2 = 4$ as a form of sensitivity analysis. We call these designs the 3-3 design, the 4-2 design and the 2-4 design, respectively.

For the new study design, our primary design used the classical approach. We set the desired power to be .8 and type I error $\alpha = .05$ level for the design. We also explore the ability to generate Bayes factors greater than 3 and 5 over the range of possible sample sizes. Since Bayes factor calculations are still somewhat computer intensive, we do not simulate the distributions of the Bayes factor at all possible values of the sample size. Instead we illustrate a simple logistic regression methodology to estimate the utility of various intermediate sample sizes.

4.2 Sampling Density for Y and Prior Density for Parameters

The sampling distribution for an n_i vector of observations Y_i for a single case indexed by i is modeled using the usual random effects model

$$\begin{aligned} Y_i &= X_i\alpha + Z_i\beta_i + \epsilon_i \\ \beta_i &\sim N_q(0, \sigma^2 D) \\ \epsilon_i &\sim N_{n_i}(0, \sigma^2 I). \end{aligned} \tag{3}$$

The design matrix X_i for a completely observed case for the original study under H_1 is 4×8, with a column of ones for the intercept, a column of zeros or ones for the effect of CS, and a 4×6 block of zeros, except for a single one in the fourth row to indicate which of the 6 TMT*CS groups the child belonged to. The Z_i matrices are n_i columns of ones in both the original and followup studies under both H_0 and H_1. Missing data within a case will cause rows of Y_i and corresponding rows of X_i and Z_i to be omitted. Under H_0, the X_i matrix is 4×2 with columns for the intercept and CS only; all columns for the treatment effect are omitted.

The pre-prior for the prior data is a flat prior, $p_0(\alpha, \sigma^2, D) \propto 1$. According to results in Hobert and Casella (1996), this prior should produce a proper posterior. Data Y_{old} from all 64 children in the original study were used with this pre-prior to produce a prior $p(\alpha, \sigma^2, D|Y_{\text{old}}, H_k)$ for $k = 0, 1$ to design the future study. We made one modification to this prior. The prior for D

was taken to be a gamma(c_0, c_1) where c_0/c_1 was set equal to the sample mean of the Gibbs sample for D from the prior, and c_0/c_1^2 was set equal to the variance of the Gibbs sample of D. Since these samples were already available from previous analyses, this required little additional work.

For the followup study, the two columns of X_i and the corresponding elements of α that refer to the N treatment are omitted. The length of Y_i will be 6 in the absence of missing data. The X_i matrix will have an initial column of ones and a second column of ones or zeros depending on the CS value. Under H_1, the third through sixth columns will be all zeros except for m_2 ones in the last m_2 rows of whichever column of X_i is the indicator variable of the CS*TMT group that the case belongs to.

Technical details of the prior and Bayes factor calculation are given in the appendix to keep this methodology self contained. With a few modifications, the prior follows the methodology in Weiss, Wang and Ibrahim (1997), hereafter WWI, with a few modifications. The data set that is used to form our prior is the data set actually analyzed in WWI. In WWI, an optional parameter may be used to calibrate the strength of the prior information contributing to the distribution of $\beta|\sigma^2, DY_{\text{old}}$. When this optional parameter is set equal to one, the prior data is contributes on an equal par with the data cases to the posterior. We set the optional parameter equal to one here. In WWI, the prior for σ^2 does not depend on D while our prior for σ^2 is directly a predictive prior for σ^2 derived from the prior data and dependent on D. Finally, our prior for D is based directly on the prior data as already described.

4.3 Covariate and Missing Data Distributions

There are two covariates in the study to be designed; CS which is binary and TMT which is binary. Of 64 children in the original study, 32 were observed to be distractors, and 32 were attenders. Starting with a uniform prior Beta$(1,1)$ for π_{CS}, the probability that a new child is a distractor, we will sample π_{CS} as a Beta$(33,33)$, and then for the n individuals in the followup study we sample CS_i as Bernoulli(π_{CS}). If the followup study were to be the same as the original, then the distribution of TMT is known to be multinomial$(1/3,1/3,1/3)$. Since we are eliminating the N treatment group, TMT is Bernoulli$(1/2)$. We use a Beta$(1,1)$ prior for all probabilities in this section. If additional information from outside the prior study data was available, the Beta prior could be altered from a Beta$(1,1)$ density.

We give three models for the missingness. The one we actually use is the third. In the original study the design called for 4 repeated measures per child for a total of $64*4=256$ observations on $n=64$ cases. However, 11 observations were missing on 6 children. To all appearances, missingness was completely at random, related to things like school absence or illness and not to an inability to follow instructions or fear or feelings about the experiment or it's results. A simple missingness model is one possibility, where $\pi_{\mathrm{miss,obs}}$

the probability that an observation is missing has prior probability density Beta$(11+1, 256-11+1)$, and every observation might be deleted at random with probability $\pi_{\text{miss,obs}}$.

A problem with this simple model is that missing observations tended to cluster; 11 observations on only 6 children is unlikely to occur by chance if observations were randomly missing. A second model for missingness models the probability of missingness $\pi_{\text{miss,case}}$ for a particular case as Bernoulli$(58+1, 6+1)$, again using a flat Beta$(1,1)$ prior. Given that a case has missing data, one observation could be deleted at random from the case, and remaining observations within the case could be deleted independently with probability π_{within} distributed a priori as Beta$(5+1, 13+1)$. This approach seems awkward and does not extend naturally to changing the number of observations within a case.

The method we actually used was to consider that children divide into two groups, *missers* and *non-missers*. Non-missers never have missing data, while each observation on a misser is missing with probability π_{within}. The probability that a child is a misser is π_{misser}. The probability that a misser has no missing data is $(1-\pi_{\text{within}})^{n_i}$, where n_i is the designed number of observations for the child. For the prior study, $n_i \equiv 4$, and for the study under design, $n_i \equiv 6$. The prior data of 58 complete data cases, and 6 missers with a total of 11 missing observations does not easily allow us to produce needed posterior samples for π_{within} and π_{misser}. However, including one extra unknown, l_0, the number of missers with zero missing observations allows us to produce a simple Gibbs sampler (Gelfand, Hills, Racine-Poon and Smith 1990) to draw samples from the posterior of π_{within} and π_{misser}. Starting from flat Beta$(1, 1)$ priors, the needed conditional distributions are

$$\begin{aligned}
\pi_{\text{within}}|\pi_{\text{misser}}, l_0, \text{prior data} &\sim \text{Beta}(11+1, 4l_0+15+1) \\
\pi_{\text{misser}}|\pi_{\text{within}}, l_0, \text{prior data} &\sim \text{Beta}(58-l_0+1, 6+l_0) \\
l_0|\pi_{\text{misser}}, \pi_{\text{within}}, \text{prior data} &\sim \text{Binomial}\left(58, \frac{\pi^4_{\text{within}}}{\pi^4_{\text{within}}+\pi_{\text{misser}}}\right)
\end{aligned}$$

The resulting mean and standard deviation of π_{misser} are .188 and .086 and for π_{within} they are .226 and .085.

4.4 Results

We simulated distributions $p(b_{01}|H_k)$ for the 3-3, 2-4 and 4-2 designs. Generally we increased n in steps of 10 starting from $n = 20$, searching for the point where the power was equal to .8. For the 3-3 design, we also investigated more carefully the power for sample sizes $n \in (37, 38, 39)$. The complete set of simulation results are reported in table 1. The first column gives the proposed sample size n, the second column is the simulation sample size, usually 100, except for $n = 38 \text{and} 39$ for the 3-3 design. Column 3 gives $\psi_{.05}(n)$, the lower 5% tail of the distribution $p(b_{01}|H_0)$ and column 4 is the power $P(b_{01} > \psi_{.05}(n)|H_1)$ which is the probability under H_1 that b_{01} is greater than the cutoff in column 3. Columns 5-7 give the probabilities that the log

Bayes factor is greater than 5, 3, and 0 if H_0 is true, and columns 8-10 give the probabilities that b_{01} is less than -5, -3, and 0 given that H_1 is true.

For the 3-3 design, the power achieves a level of .80 for a sample of size $n = 39$. We might expect the cutoff points $\psi_{.05}(n)$ to be monotone in n, however, they are not. This is because of sampling variability in the calculations. The standard error of estimation of $\psi_{.05}(n)$ from a sample of size 100 is approximately .9, and for a sample of size 1000, the standard error is approximately .3, so none of the inversions or equalities are too surprising. For the 2-4 design we appear to need less than 40 observations while for the 4-2 design we will need somewhat over 50 observations. The 4-2 design does appear to have less power than the other two, while the 3-3 and 2-4 are close in power.

Inspection of table 1 suggests that if we wanted to make $P(b_{01} > 3|H_0) + P(b_{01} < -3|H_1) \geq 1$, then the necessary sample size appears to be barely over 20 for the 3-3 and 2-4 designs, but will be close to 30 for the 4-2 design. If we try for $P(b_{01} > 5|H_0) + P(b_{01} < -5|H_1) \geq 1$, then we need an n of slightly over 50 for the 3-3 design, slightly under 50 for the 2-4 design, and approximately 60 for the 4-2 design. We will get more accurate conclusions shortly.

Since the simulation sample sizes in table 1 are not as large as we might desire, and simulations could not be run at all sample sizes, it is helpful to borrow strength from different simulations to estimate the design characteristics for different sample sizes. We did this by fitting a logistic regression model to each of the 7 columns in each of the 3 sections of the table, for a total of 21 logistic regressions. As an example, for the 3-3 design, and for the probability $P(b_{01} > 3|H_0)$, which is the 6^{th} column of the top portion of the table, there are 7 data points, with predictor values $n = (20, 30, 37, 38, 39, 40, 50)$, response y equal to the proportion of times that $b_{01} > 3$ – this is $y = 67$ for $n = 20$ and $y = 71$ for $n = 30$, for example. Then $y \sim \text{binomial}(L, \pi)$ with L equal to the number of simulations, usually 100, but for $n = 38, 39$, $L = 1200$ and 800; and $\log(\pi/(1-\pi)) = \beta_0 + \beta_1 n$. Since the probabilities within a column do not vary greatly, a linear logistic regression can be expected to do a good job of interpolating and smoothing the results of the study. Table 2 has the same format as columns 1 and 4-10 of table 1 but the tabled probabilities are fitted results from these logistic regressions.

Fitting these logistic regressions and using the resulting estimated values to re-estimate the probabilities in table 1 substantially increases the effective sample size of most results. For example, for the 3-3 design and a sample size of 40, using the fitted results to estimate $P(b_{01} < -3|H_1)$ gives an estimate of .80 with a standard error of .02 and an effective sample size of 325 rather than the simulation sample size of 100 and simulation standard error of approximately .04 and a simulation estimate of .88 which happened to be the same as the simulation result for $n = 50$. Furthermore, we can estimate this probability for $n = 41$, giving an estimated probability of .806 and a similar effective sample size in spite of not having run simulations for $n = 41$.

Interpolation of table 2 suggests that we need 39 cases for the 3-3 design to have power of .8, 34 cases for the 2-4 design and fully 53 cases for the 4-2 design. The smoothing makes these results more dependable than the sample size estimates based on table 1. The sample sizes needed to produce $P(b_{01} > 3|H_0) + P(b_{01} < -3|H_1) \geq 1$ are 21, 23, and 32 for the 3-3, 2-4, and 4-2 designs. And sample sizes to give $P(b_{01} > 5|H_0) + P(b_{01} < -5|H_1) \geq 1$ are 50, 48, and 60.

The calculations $P(b_{01} > 0|H_0)$ and $P(b_{01} < -0|H_1)$ are of interest in sample size specification since minimally we would like a study design that has the Bayes factor of the correct sign, either positive or negative given that H_0 or H_1 is true. Under H_0, we see that there is always substantial probability of at least .89 that b_{01} is greater than 0 for all sample sizes. Under H_1, the probabilities are lower of having the correct sign, ranging from .61 to .80. The reason that the power is higher than $P(b_{01} < 0|H_1)$ is that the cutoff is positive at the higher values of n so that $P(b_{01} < \psi_{.05}(n)|H_1) > P(b_{01} < 0|H_1)$.

Generally we expected the 2-4 design to be the most efficient, since extra observations are taken after treatment when there are four groups and fewer observations are taken before treatment when there are only two groups. Similarly, we expect the 3-3 design to be more efficient than the 4-2 design however this was not strongly shown in investigations of table 2. Generally the results follow the expectation, although occasionally at the lowest sample sizes, $n = 20$, and sometimes $n = 30$, the 4-2 design can beat the 3-3 design and even the 2-4 design.

Interestingly, the slopes of the logistic regressions that produced table 2 are identical to within sampling error and with a few exceptions such as for the 3-3 design and $P(b_{01} < 0|H_1)$ where the probabilities are all quite high. For example, the slopes for the 3-3 design center around .035 except for power which has a slope of .094, and $P(b_{01} < 0|H_1)$ which has a slope of .017. This suggests that further combining of the simulation results might be possible to increase accuracy; however this lead was not followed here.

Approximately sixty children in the 8-10 age range are available for the followup study. As with most study design, sample size selection is driven not only by power but by cost, subject availability and other considerations. The analysis here shows that even for the least efficient 4-2 design and $n = 60$ subjects, we have (a) sufficient power and (b) reasonable probability of determining which hypothesis is correct using a Bayes factor approach.

5 Discussion

The priors used for calculating b_{01} could be different from the data generation priors $p(\theta|H_j)$ but this requires explicit justification. A reason for considering multiple priors would be sensitivity analysis; an important class of sensitivity analyses would be to assess how other informed interests such as colleagues and funding agencies assess any proposed sample size. However, these analyses would seem to require that the data generation and prior distribution densities be the same. Some experience (see WWI for one example) with repeated measures random effects priors suggest that an informative prior is helpful for producing a large Bayes factor. Obviously an informative prior requires justification; taking an informative prior merely to produce a large Bayes factor is not appropriate. Prior data, as in our example, seems the best way of producing such an informative prior.

References

Chen, M. H., Ibrahim, J. G. and Yiannoutsos, C. (1996). Prior Elicitation and Bayesian Computation for Logistic Regression Models with Applications to Variable Selection. Harvard School of Public Health Department of Biostatistics technical report.

Chen, M. H, and Shao, Q. M. (1997a). Estimating Ratios of Normalizing Constants for Densities with Different Dimensions. *Statistica Sinica, 7*, 607-630.

Chen, M. H. and Shao, Q. M. (1997b). On Monte Carlo Methods for Estimating Ratios of Normalizing Constants. *Annals of Statistics*, 25, to appear.

Chib, S. (1995). Marginal Likelihood From the Gibbs Output. *Journal of the American Statistical Association*, 90, 1313-1321.

Fanurik, D., Zeltzer, L. K., Roberts, M. C., and Blount, R. L. (1993). The Relationship Between Children's Coping Styles and Psychological Interventions for Cold Pressor Pain. *Pain*, 53, 213-222.

Freedman, L. S. and Spiegelhalter, D. J. (1983). The assessment of subjective opinion and its use in relation to stopping rules for clinical trials. *The Statistician*, 32, 153-160.

Geyer, C. (1994). Estimating Normalizing Constants and Reweighting Mixtures in Markov Chain Monte Carlo. Technical report 568, School of Statistics, University of Minnesota.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association*, 85, 972-985.

Gelman, A. and Meng, X. (1994). Path Sampling for Computing Normalizing Constants: Identities and Theory. Technical Report 377, Department of Statistics, University of Chicago.

Hobert, J. P. and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *Journal of the American Statistical Association*, 91, 1461-1473.

Kass, R. E. and Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773-795.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data.* New York: Wiley.

Lui, K.-J. and Cumberland, W. G. (1992). Sample Size Requirements for Repeated Measurements in Continuous Data. *Statistics in Medicine*, 11, 633-641.

Meng, X. and Wong, W. H. (1994) Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration. Statistica Sinica, to appear.

Muller, K. E., LaVange, L. M., Ramey, S. L. and Ramey, C. T. (1992). Power Calculations for General Linear Multivariate Models Including Repeated Measures Applications. *Journal of the American Statistical Association*, 87, 1209-1226.

Müller, P. and Parmigiani, G. (1995). Optimal Design via Curve Fitting of Monte Carlo Experiments. *Journal of the American Statstical Association*, 90, 1322-1330.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society*, Ser. B **56**, 3-48.

Rochon, J. (1991). Sample Size Calculations for Two-Group Repeated-Measures Experiments. *Biometrics*, 47, 1383-1398.

Scott, D. (1992). Multivariate Density Estimation. New York: Wiley.

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Statistics in Medicine*, 5, 1-13.

Verdinelli, I. (1996). *Bayesian Design of Experiments for the Linear Model.* Carnegie Mellon University, Unpublished PhD Thesis.

Weiss, R. E. (1994). Pediatric Pain, Predictive Inference and Sensitivity Analysis. *Evaluation Review*, 18, 651-678.

Weiss, R. E. (1996). Sufficiency and Influence (with Discussion). In Bayesian Robustness, IMS Lecture Notes, Monograph Series, Volume 29, 213-230.

Weiss, R. E. (1997). Bayesian Sample Size Calculations for Hypothesis Testing. *The Statistician*, 46, 185-191.

Weiss, R. E., Wang, Y. and Ibrahim, J. G. (1997). Predictive Model Selection for Repeated Measures Random Effects Models Using Bayes Factors. *Biometrics*, 53, 159-169.

A Prior Density and Bayes Factor Calculation

Here we give the exact form of the prior density and technical details about the Bayes factor calculation.

Let $x \sim \mathrm{IG}(a, b)$ represent an inverse gamma distributed random variable with density

$$\frac{b^a}{x^{a+1}\Gamma(a)} \exp\left(-\frac{b}{x}\right);$$

with $\Gamma(a)$ being the usual Gamma function; let $x \sim N(a, b)$ be a normally distributed random variable with mean a and variance b; and let $x \sim G(a, b)$ denote a gamma distributed random variable with density proportional to

$$\frac{b^a x^{a-1}}{\Gamma(a)} \exp(-bx);$$

The model for the prior data and for the study to be designed is given by (3). For a generic data set modeled using (3), define $Y = (Y_1^t, \ldots, Y_n^t)^t$, $X = (X_1^t, \ldots, X_n^t)^t$, and $Z = \mathrm{diag}(Z_1, \ldots, Z_n)$. Let $N = \sum_{i=1}^n n_i$ be the total number of observations taken on all n people in the study and let p be the number of columns of the X matrix, which is $p_{\mathrm{old}} = 8$ under H_1 in the old study and $p_{\mathrm{new}} = 6$ in the new, since we lose two columns in the new study by design and $p_{\text{old}} = p_{\text{new}} = 2$ under H_0. References to either the old or the new data will be subscripted by the words old or new as Y_{old}, X_{old}, Z_{old}, p_{old} or N_{old} for example. To reduce clutter, a subscript $k = 0$ or 1 for the model will not be used and dependence of densities on H_k will be avoided except for $f(Y_{\text{new}}|Y_{\text{old}}, H_k)$. The calculations described here lead to $f(Y_{\text{new}}|Y_{\text{old}}, H_k)$ and so need to be performed once under each H_k to calculate the Bayes factor.

Formally, the prior $p(\alpha, \sigma^2, D|Y_{\text{old}}) = p(\alpha|\sigma^2, D, Y_{\text{old}})p(\sigma^2|D, Y_{\text{old}})p(D|Y_{\text{old}})$ is

$$
\begin{aligned}
\alpha|\sigma^2, D, Y_{\text{old}} &\sim \mathrm{N}(\alpha_0, \sigma^2 A) \\
\sigma^2|D, Y_{\text{old}} &\sim \mathrm{IG}\left(\frac{N_{\text{old}} - p_{\text{old}} + 2}{2}, \frac{\mathrm{RSS}_{\text{old}}(D)}{2}\right) \\
D|Y_{\text{old}} &\sim \mathrm{gamma}(c_0, c_1)\,,
\end{aligned}
$$

where c_0 and c_1 are constants derived earlier as explained in subsection 4.2, and

$$
\begin{aligned}
\alpha_0 &= (X_{\text{old}}^t V_{\text{old}}^{-1} X_{\text{old}})^{-1} X_{\text{old}}^t V^{-1} Y_{\text{old}} \\
V_{\text{old}} &= I + Z_{\text{old}}(I \otimes D) Z_{\text{old}}^t \\
A &= (X_{\text{old}}^t V_{\text{old}}^{-1} X_{\text{old}})^{-1} \\
\mathrm{RSS}(D) &= Y^t Q_Z (I - P_X) Y \\
Q_Z &= I - Z(Z^t Z + I \otimes D) Z^t \\
P_X &= X(X^t Q_Z X)^{-1} X^t Q_Z\,.
\end{aligned}
$$

We need to calculate the prior predictive sampling density $f(Y_{\text{new}}|Y_{\text{old}}, H_k)$

for the numerator, $k = 0$ and denominator $k = 1$ of the Bayes factor B_{01}

$$f(Y_{\text{new}}|Y_{\text{old}}, H_k) = \int f(Y_{\text{new}}|\alpha, \sigma^2, D)p(\alpha, \sigma^2, D|Y_{\text{old}})d\,\alpha\,d\,\sigma^2\,d\,D\,.$$

The inner two integrals can done in closed form giving

$$f(Y_{\text{new}}|Y_{\text{old}}, H_k) = \int f(Y_{\text{new}}, D|Y_{\text{old}})\,d\,D$$

where

$$f(Y_{\text{new}}, D|Y_{\text{old}}) = \frac{(.5\text{RSS}_{\text{old}}(D))^{.5(N_{\text{old}}-p_{\text{old}}+2)}|V_{\alpha\alpha}|^{1/2}|Q_{Z_{\text{new}}}|^{1/2}\Gamma(.5(N_{\text{new}}+N_{\text{old}}-p_{\text{old}}+2))p(D|Y_{\text{old}})}{(2\pi)^{N_{\text{new}}/2}|A|^{1/2}(.5M(Y_{\text{new}}, D))^{.5(N_{\text{new}}+N_{\text{old}}-p_old+2)}\Gamma(.5(N_{\text{old}}-p_{\text{old}}+2))},$$

where

$$\begin{aligned}
V_{\alpha\alpha} &= X^tX + A^{-1} \\
M(Y_{\text{new}}, D) &= (Y_{\text{new}} - X\alpha^*)^t Q_{Z_{\text{new}}}(Y_{\text{new}} - X\alpha^*) \\
&\qquad +(\alpha^* - \alpha_0)^t A^{-1} V_{\alpha\alpha} X^t Q_{Z_{new}} X(\alpha^* - \alpha_0) + \text{RSS}_{\text{old}}(D) \\
\alpha^* &= (X^t_{\text{new}} Q_{Z_{\text{new}}} X)^{-1} X^t Q_{Z_{\text{new}}} Y_{\text{new}}
\end{aligned}$$

This last integral with respect to D we do numerically using Simpson's or other one dimensional integration rule.

3-3 design

n	L	$\psi_{.05}(n)$	power	$P(b_{01} > a\|H_0)$ a=5	a=3	a=0	$P(b_{01} < -a\|H_1)$ a=5	a=3	a=0
20	100	-2.00	0.38	0.32	0.67	0.89	0.20	0.31	0.61
30	100	-1.00	0.60	0.31	0.71	0.91	0.30	0.40	0.63
37	100	0.40	0.76	0.49	0.81	0.97	0.30	0.41	0.71
38	1200	0.75	0.79	0.43	0.78	0.97	0.39	0.53	0.74
39	800	0.50	0.80	0.43	0.79	0.97	0.37	0.52	0.75
40	100	1.13	0.84	0.47	0.88	0.99	0.36	0.47	0.75
50	100	1.50	0.86	0.56	0.88	0.97	0.39	0.53	0.76

2-4 design

n	L	$\psi_{.05}(n)$	power	$P(b_{01} > a\|H_0)$ a=5	a=3	a=0	$P(b_{01} < -a\|H_1)$ a=5	a=3	a=0
20	100	-0.05	0.63	0.23	0.67	0.94	0.19	0.28	0.64
30	100	0.70	0.72	0.31	0.75	0.96	0.29	0.45	0.67
40	100	0.70	0.81	0.39	0.80	0.97	0.33	0.45	0.72
50	100	2.80	0.93	0.63	0.92	1.00	0.48	0.60	0.83

4-2 design

n	L	$\psi_{.05}(n)$	power	$P(b_{01} > a\|H_0)$ a=5	a=3	a=0	$P(b_{01} < -a\|H_1)$ a=5	a=3	a=0
20	100	0.10	0.68	0.24	0.56	0.95	0.10	0.25	0.65
30	100	-0.50	0.62	0.27	0.68	0.93	0.25	0.35	0.65
40	100	0.50	0.77	0.34	0.67	0.97	0.27	0.47	0.74
50	100	0.40	0.79	0.38	0.77	0.97	0.44	0.53	0.76
60	100	1.30	0.83	0.57	0.81	0.98	0.42	0.52	0.72

Table 1: Top, middle, and bottom are for the 3-3, 2-4 and 4-2 designs respectively. Column 1: proposed sample size. Column 2: Simulation sample size L. Column 3: $\psi_{.05}(n)$ is the fifth percentile of the distribution of the log Bayes factor b_{01} under H_0. Column 4: Power is $P(b_{01} < \psi_{.05}(n)|H_1)$ the probability under H_1 that b_{01} is less than the cutoff in column 3. Columns 5-7: Probability that b_{01} is greater than 5, 3 or 0 under H_0. Columns 8-10: Probability that b_{01} is less than -5, -3 or 0 under H_1.

3-3 design

n	power	$P(b_{01} > a \mid H_0)$ a=5	a=3	a=0	$P(b_{01} < -a \mid H_1)$ a=5	a=3	a=0
20	.40	.29	.64	.89	.25	.35	.61
30	.63	.36	.73	.94	.31	.44	.68
37	.77	.42	.78	.97	.36	.50	.73
38	.78	.43	.79	.97	.37	.51	.74
39	.80	.44	.79	.97	.38	.52	.74
40	.81	.45	.80	.97	.38	.53	.75
50	.92	.54	.86	.99	.46	.61	.80

2-4 design

n	power	$P(b_{01} > a \mid H_0)$ a=5	a=3	a=0	$P(b_{01} < -a \mid H_1)$ a=5	a=3	a=0
20	.61	.21	.65	.93	.19	.30	.61
30	.74	.32	.76	.96	.27	.39	.67
40	.84	.45	.84	.98	.36	.49	.72
50	.90	.59	.89	.99	.47	.59	.77

4-2 design

n	power	$P(b_{01} > a \mid H_0)$ a=5	a=3	a=0	$P(b_{01} < -a \mid H_1)$ a=5	a=3	a=0
20	.64	.21	.57	.94	.14	.28	.65
30	.69	.28	.64	.95	.20	.35	.68
40	.74	.35	.70	.96	.28	.42	.71
50	.79	.43	.76	.97	.37	.50	.73
60	.83	.52	.81	.98	.48	.57	.75

Table 2: Fitted values. Table is smoothed estimates of the probabilities from table 1. Column 1: Proposed sample size. Column 2: Power is $P(b_{01} < \psi_{.05}(n) \mid H_1)$. Columns 3-5: Probability that b_{01} is greater than 5, 3 or 0 under H_0. Columns 6-8: Probability that b_{01} is less than -5, -3 or 0 under H_1.