# A Bayesian sample size determination method with practical applications

S. K. Sahu and T. M. F. Smith†

*University of Southampton, UK*

**Summary**. The problem motivating this article is the determination of sample size in clinical trials under normal likelihoods and at the substantive testing stage of a financial audit where normality is not an appropriate assumption. A combination of analytical and simulation based techniques within the Bayesian framework is proposed. The framework accommodates two different prior distributions: one is the general purpose fitting prior distribution used in Bayesian analysis and the other is the expert subjective prior distribution, the sampling prior which is believed to generate the parameter values which in turn generate the data. We obtain many theoretical results and one key result is that typical non-informative prior distributions lead to very small sample sizes. On the other hand, a very informative prior distribution may either lead to a very small or a very large sample size depending on the location of the centre of the prior distribution and the hypothesized value of the parameter. The methods developed here are quite general and can be applied to other sample size determination (SSD) problems. A number of numerical illustrations which bring out many other aspects of the optimum sample size are given.

*Keywords*: Auditing; Bayesian inference; book values; clinical trials; fitting prior; mixture distribution; rare errors; simulation based approach; sampling prior; taints.

## 1. Introduction

The problem motivating this paper is sample size determination (SSD). It arose in two areas: clinical trials in medicine and substantive tests in auditing. In clinical trials the SSD is a well debated problem and Chapter 6 of the book by Spiegelhalter *et al.* (2004) and the references therein provide an excellent overview of current issues. The difficulties that arise are related to the practical issues of medical relevance of the specification of null and alternative hypotheses and the choice of fixed error rates for size and power, see for example, Spiegelhalter *et al.* (1994). The optimal sample size in the classical framework depends crucially on the choice of the alternative hypothesis. Spiegelhalter and Freedman (1986) argue that often such a choice is dependent on 'bewildering and rather ill-defined recommendations'. This is not the case in a Bayesian formulation of the problem since there is no need to specify a particular value of the alternative hypothesis.

Financial auditing involves several stages. At the first stage senior auditors review the system generating the accounts and compare the current results with those of previous years and with those of similar entities. In the light of this review a strategy for more detailed explorations and tests is developed. The next stage is to test the working of the

†*Address for correspondence:* Sujit K. Sahu, School of Mathematics, $S^3$RI, University of Southampton, Southampton, SO17 1BJ, United Kingdom
E-mail: S.K.Sahu@maths.soton.ac.uk

accounting system and, in particular, the implementation of controls and checks. This phase is known as compliance testing and may exceptionally be done using a computer generated set of transactions, running them through the system and checking for compliance. The *substantive testing* of actual transactions follows. Errors in money values are rarely found in samples selected from well-designed accounting systems and it is this paucity of actual values of errors that makes the SSD problem so difficult.

There is a considerable literature on the analysis of audit data, see e.g. Smith (1976, 1979), Laws and O'Hagan (2000, 2002) and the references therein. The information from the early stages of an audit is mainly qualitative and often leads to strong opinions about the quality of the system. Combining this prior information with the hard data generated by sampling at the substantive stage may be done in an ad-hoc manner within the frequentist tradition, see e.g. Heiner and Whitby (1980), Patterson (1993), and Shrivastava and Shafer (1994), or more formally using Bayes' theorem. An important reference is Cox and Snell (1979) who propose a Bayesian mixture model for the analysis of substantive data. See Laws and O'Hagan (2000, 2002) for an extension of this model. The practical problem is that if money errors are rare then the number of errors found in small or medium sized samples will be very small, and possibly zero. Thus the effective sample size for frequentist inference about the total of money errors is small and the resulting inferences will be unreliable. Using the available prior information within a Bayesian methodology should lead to more reliable conclusions about the unknown error totals. Here standard frequentist methods based on normal approximations are not appropriate and the alternative solutions proposed have often been rejected by auditors since they give sample sizes and error limits that are far larger than their expectations.

For SSD in any area the only information available is prior information. Introducing uncertainty into prior estimates is a quintessentially Bayesian procedure and so we explore the use of Bayesian methods for SSD within distributional frameworks relevant to auditing and clinical trials. In both cases there are practical constraints that require the sample sizes to be determined in advance, and so we assume that the objective is to determine an optimal fixed sample size that satisfies a criterion based on the Bayes' risk. Given specific loss functions and sampling cost functions it is possible to carry out a full Bayesian analysis for SSD, see Raiffa and Schlaifer (2000), Lindley (1997) and the references therein. In the absence of precise information about costs and losses we approximate the loss functions and employ an approximate Bayesian approach in the spirit of Adcock (1997), Joseph *et al.* (1995), and Wang and Gelfand (2002), that should give reasonable estimates of sample size.

In this article we adopt the framework proposed in Wang and Gelfand (2002) where two different prior distributions are used for SSD. They propose that the prior for inference, the fitting prior, can differ from the prior used for averaging in the calculation of the Bayes risk, the sampling prior. Spiegelhalter *et al.* (2004) also propose the use of two different priors, an enthusiastic and a sceptical prior, for the analysis and monitoring of clinical trials. However, for SSD they propose a hybrid Bayes/frequentist approach using a single prior elicited from a team of experts. We shall see that the fitting prior distribution does not influence the sample size much if it is assumed to be non-informative. The sampling prior, on the other hand, has a large influence on the optimal sample size. We explore these and other consequences of using different fitting and sampling priors for SSD in our simulation studies in Sections 3 and 4.

The plan of the remainder of this article is as follows. In Section 2 we develop the general methodology. In Section 3 we discuss the problem of SSD in clinical trials, and

present results for the normal distribution with an example. In Section 4 we discuss the problem of SSD in auditing and present results for a mixture distribution first proposed by Cox and Snell (1979) for this problem. We then illustrate these results numerically. The article ends with some summary remarks in Section 5. The derivations of the theoretical results are presented in Appendices.

## 2. Method

In both the auditing and clinical trial problems the objective of the data analysis is to test a specific hypothesis. The particular hypotheses are described more fully in Sections 3 and 4. In each case the problem is to choose a sample size while taking into account the consequences of wrong decisions about the hypothesis under test. To choose a sample size is to make a decision. A full Bayesian approach to decision making requires the specification of probability distributions for both the data and the unknown parameters, a list of possible actions, the losses consequent on wrong actions, and the cost of sampling. In the absence of any of these components approximations must be made to the full Bayesian approach, and this is the line that we take in preference to abandoning the full Bayesian approach.

### *2.1. The hypothesis testing problem*

Let $\mathbf{X}^{(n)} = (X_1, \ldots, X_n)$ denote a random sample of size $n$ from a population with density $f(x|\theta)$ and let $\pi(\theta)$ denote the prior distribution for the unknown parameter $\theta$. Let $\pi(\theta|\mathbf{x}^{(n)})$ denote the posterior distribution of $\theta$ given the observed sample $\mathbf{x}^{(n)}$.

We follow the development in Berger (1985, Chapter 7) to set up the hypothesis testing problem which is to choose between the two hypotheses:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1,$$

where $\Theta_0$ is less than $\Theta_1$ in the sense that, if $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ then $\theta_0 < \theta_1$. In this article we shall take $\Theta_0 = \{\theta : -\infty < \theta \leq \theta_0\}$ and $\Theta_1 = \{\theta : \theta_0 < \theta < \infty\}$.

In clinical trials $\theta_0$ is the null value, possibly zero, of the difference between a treatment and a control. The choice of this null value is controversial, Spiegelhalter and Freedman (1986), and is the responsibility of the clinicians. We discuss this further in Section 3.

In the auditing context $\theta_0$ represents a value corresponding to a material error per item. If $\theta < \theta_0$ the error is not material and the account will be accepted. If, however, $\theta \geq \theta_0$ the error is material and the account will be rejected and the auditors will qualify that section of the accounts in their conclusions. Note that $\theta_0$ *is a positive quantity* set in advance by the auditors, not by statisticians. Setting $\theta = \theta_0$ as a null hypothesis is not sensible for the auditing problem as the critical region will fall into an area corresponding to a material error.

Let $a_i$ denote the action of accepting $H_i$ for $i = 0, 1$ and $L(\theta, a_i)$ denote the loss for taking decision $a_i$ when $\theta$ is the true value. The Bayes decision rule, denoted by $\delta_n^\pi$, is to select $a_0$ if the average posterior loss under $a_0$ is less than that under $a_1$, that is, if

$$\int_{\Theta_1} L(\theta, a_0)\pi(\theta|\mathbf{x}^{(n)})d\theta < \int_{\Theta_0} L(\theta, a_1)\pi(\theta|\mathbf{x}^{(n)})d\theta. \tag{1}$$

Under some parametric assumptions it is often possible to find a suitable function $g\left(\mathbf{x}^{(n)}\right)$ such that (1) holds if and only if $g\left(\mathbf{x}^{(n)}\right) < k^\pi(n)$ where $k^\pi(n)$ is the value of $g\left(\mathbf{x}^{(n)}\right)$ for

which equality holds in (1) instead of the inequality. In the parametric family $f(x|\theta)$, if $\bar{X}_n$ is sufficient for $\theta$ then Berger (1985) establishes that $g\left(\mathbf{x}^{(n)}\right) = \bar{x}_n$. This will be the case for our normal error distribution in Section 3.1. However, this simplification is not possible for our mixture model and in Section 4.1 we work with the appropriate $g$ function.

The Bayes decision risk prior to sampling, denoted by $r\left(\pi, \delta_n^{\pi}\right)$, is given by

$$\begin{aligned} r(\pi, \delta_n^{\pi}) &= \int_{\Theta_1} L(\theta, a_0) P\left\{g\left(\mathbf{X}^{(n)}\right) < k^{\pi}(n)|\theta\right\} \pi(\theta) d\theta \\ &\quad + \int_{\Theta_0} L(\theta, a_1) P\left\{g\left(\mathbf{X}^{(n)}\right) \geq k^{\pi}(n)|\theta\right\} \pi(\theta) d\theta, \end{aligned} \tag{2}$$

where $P(\cdot|\theta)$ denotes the probability of its argument when $\theta$ is the true value.

For SSD we may also define a cost function, $c(n)$ say, for obtaining the samples. In general SSD problems $c(n)$ is often chosen to be an increasing function of $n$ which does not involve the parameters in the likelihood or prior distribution, while the risk decreases with $n$. The SSD problem is to minimise

$$r(\pi, \delta_n^{\pi}) + c(n)$$

over the values of the sample size, $n$. The smallest $n$ which minimises the above is the required sample size. In view of this we reformulate the SSD problem for an unknown cost function as that of bounding the risk $r\left(\pi^{(s)}, \delta_n^{\pi^{(f)}}\right)$ by a pre-specified quantity.

Note that the parametric assumption enters the sample size calculation through the probability $P\left\{g\left(\mathbf{X}^{(n)}\right) \geq k^{\pi}(n)|\theta\right\}$. The non-parametric approach of Walker (2003) approximates this probability using the central limit theorem. Therefore, the optimum sample sizes for the Gaussian model should be similar to the ones obtained from an equivalent non-parametric approach for large sample sizes.

### 2.2. Fitting and sampling priors

All Bayesian model fitting exercises need a prior distribution for the unknown parameters in the model. This is the prior distribution which would have been used for model fitting if the sample data were available. Following Wang and Gelfand (2002) we call this the fitting prior and denote it by $\pi^{(f)}(\theta)$. Often, $\pi^{(f)}(\theta)$ is assumed to be vague (or non-informative) so that the modeller encourages the data to drive the inference, thus it is a general purpose working prior distribution.

The fitting prior is to be used to obtain the posterior distribution $\pi(\theta|\mathbf{x}^{(n)})$ in (1) and to emphasise this dependence we write the posterior distribution as $\pi^{(f)}(\theta|\mathbf{x}^{(n)})$. Thus the decision rule is denoted by $\delta_n^{\pi^{(f)}}$ and it selects $a_0$ if (1) holds for the posterior distribution $\pi^{(f)}(\theta|\mathbf{x}^{(n)})$. The quantity $k^{\pi}(n)$ will also depend on the fitting prior used to calculate the posterior distribution and we emphasise this dependence by writing $k^{\pi^{(f)}}(n)$.

In the frequentist approach to the SSD problems it is usually of interest to investigate the sensitivity of the SSD procedure when the 'true' parameter $\theta$ assumes some particular values. This is not considered to be satisfactory from a Bayesian perspective where the unknown parameter $\theta$ is assumed to be random. To perform sensitivity analysis in a coherent Bayesian framework it is natural to assume that the parameter $\theta$ follows an informative prior distribution concentrated around some specific values of $\theta$ which are of particular interest to the practitioner. This is the prior that a pure Bayesian would employ after full consideration

of all the available prior information. Wang and Gelfand (2002) formalised this concept by calling this informative prior distribution the sampling prior. Here this prior is denoted by $\pi^{(s)}(\theta)$ and it replaces the familiar assumption of fixing $\theta$ in the classical SSD problem.

*What are the differences between the fitting and sampling priors and why should they not be the same?* The sampling prior is the prior distribution used to generate the parameter values which are then conditioned upon to generate the data from $f(x|\theta)$ in substantive experiments. That is, data $\mathbf{X}^{(n)}$ are generated from the joint hierarchical model $\pi^{(s)}(\theta)f(x|\theta)$. Once data are available we would like to pretend that the informative prior distribution which generated the data is unknown to us; and we would like to make inference with the assumption of a relatively non-informative prior distribution. The sampling and fitting prior distributions should not be the same because they serve two different purposes in the SSD problems. The sampling prior distribution addresses the 'what if' type sensitivity scenarios, whereas the fitting prior distribution is used to form the posterior distribution for making inference. In our numerical illustrations we will investigate the situation where the sampling prior is the same as the fitting prior, the conventional Bayesian approach, and also explore the effect of different sampling and fitting priors.

The distinction between the sampling and fitting prior distributions will naturally affect the calculation of the Bayes risk, $r(\pi, \delta_n^{\pi})$ given in (2). As mentioned above the decision rule $\delta_n^{\pi}$ will need to be written as $\delta_n^{\pi^{(f)}}$. The prior distribution $\pi(\theta)$, used as the averaging measure in the integrals of (2), will be the sampling prior distribution $\pi^{(s)}(\theta)$. Thus the Bayes risk (2) will have the following form:

$$\begin{aligned} r\left(\pi^{(s)}, \delta_n^{\pi^{(f)}}\right) &= \int_{\Theta_1} L(\theta, a_0) P\left\{g\left(\mathbf{X}^{(n)}\right) < k^{\pi^{(f)}}(n)|\theta\right\} \pi^{(s)}(\theta) d\theta \\ &\quad + \int_{\Theta_0} L(\theta, a_1) P\left\{g\left(\mathbf{X}^{(n)}\right) \geq k^{\pi^{(f)}}(n)|\theta\right\} \pi^{(s)}(\theta) d\theta. \end{aligned} \tag{3}$$

### 2.3. Specific losses and bounding the risk

In typical SSD problems a specific loss function needs to be assumed. There is a general consensus that the loss function is a bounded function taking the value zero if a correct decision is made. Often practitioners are very reluctant to specify a particular function or absolute values of losses. However, we have found that they feel more comfortable in specifying the ratio of losses defined below. Assuming the constant loss function $L(\theta, a_0) = L_0$ for $\theta > \theta_0$ and $L(\theta, a_1) = L_1$ for $\theta \leq \theta_0$, practitioners may provide the ratio of losses, $L_0/L_1$, or equivalently

$$\eta = \frac{L_0}{L_0 + L_1}.$$

Henceforth, we shall work with this particular loss function and the ratio wherever possible, although the methodology can be applied more generally. Even with this assumption of constant loss function we shall see in Sections 3 and 4 that it is not possible to obtain the exact analytical sample size. However, below we obtain an attractive interpretation of risk in terms of two error probabilities and in Section 3 we obtain asymptotic results in terms of the prior sample size.

Now we have the following simple form of the risk function (3)

$$\begin{aligned} r\left(\pi^{(s)}, \delta_n^{\pi^{(f)}}\right) &= L_0 \left[\int_{\Theta_1} P\left\{g\left(\mathbf{X}^{(n)}\right) < k^{\pi^{(f)}}(n)|\theta\right\} \pi^{(s)}(\theta) d\theta \right. \\ &\quad \left. + \tfrac{1-\eta}{\eta} \int_{\Theta_0} P\left\{g\left(\mathbf{X}^{(n)}\right) \geq k^{\pi^{(f)}}(n)|\theta\right\} \pi^{(s)}(\theta) d\theta\right]. \end{aligned}$$

The above risk function is a multiple of the loss $L_0$ and it depends on the ratio of the losses $\eta$. In the absence of the absolute values of the losses we re-formulate the SSD problem as one of finding the minimum $n$ such that

$$\frac{r\left(\pi^{(s)}, \delta_n^{\pi^{(f)}}\right)}{L_0} \leq M(\eta)$$

for given values of $\eta$ and $M(\eta)$. Note that this is a canonical version of the SSD problem which bounds the risk by $L_0 M(\eta)$. Also under the assumption that $L_0 = L_1$, i.e. the losses are equal for the two possible wrong decisions, we see that the quantity to be bounded for the SSD is the sum of two error probabilities, which is an appealing quantity to bound for practical problems. In our numerical illustrations we shall experiment with three values of $M(\eta)$, viz. 0.25, 0.15 and 0.10. These particular values can be interpreted as: the test of $H_0$ is carried out at 5% level of significance and it is required to have 80%, 90% and 95% power, respectively. Of course, the last one implies a very strict condition on the two error probabilities and we shall see that many sample sizes will be very large. We set the optimum sample size to be $\infty$ if it is greater than 5000.

## 3. Application in clinical trials

The SSD problem in designing clinical trials to compare a new treatment against a standard one has received a lot of attention in the literature. See, for example, Spiegelhalter *et al.* (2004, Chapter 6) for a recent review. They discuss a range of issues including the differences between classical and Bayesian methods, use of loss functions, specification of null hypotheses, and ethical considerations for randomisation. In this paper we do not re-visit those discussions, rather our goal is to apply the methodology of Section 2 to the SSD problem.

For many SSD problems in clinical trials and elsewhere, the likelihood is approximated by a normal distribution by appealing to the central limit theorem for a summary statistic such as the log odds ratio, see for example Spiegelhalter *et al.* (2004, Section 2.4). Even when using a non-parametric model the central limit theorem may be used to approximate some key probabilities required for the SSD problem, see e.g. Walker (2003), Clarke and Yuan (2002) and Section 2.1 for more in this regard. In the remainder of this section we assume that the observables, the $X_i$'s, are normally distributed. As is expected, this turns out to be an analytically tractable situation where our methods provide some exact solutions, though the final sample size still needs to be calculated using computer intensive methods.

### 3.1. Sample size under normal likelihoods

Suppose that $X|\theta \sim N(\theta, \sigma^2)$ where $\sigma^2$ is known and assume $\pi^{(f)}(\theta) = N(\mu_f, \tau_f^2)$ and $\pi^{(s)}(\theta) = N(\mu_s, \tau_s^2)$. All hyper-parameters are assumed to be known.

Using the derivations in Appendix A under the assumptions in Section 2.3 we investigate the risk function and obtain analytical solutions. The risk function as given in (9) is:

$$r(\pi^{(s)}, \delta_n^{\pi^{(f)}}) = L_0 P\left(U > a, V < b\right) + L_1 P\left(U < a, V > b\right),$$

where $U$ and $V$ jointly follow the bivariate normal distribution with zero means, unit variances and correlation $\rho$, and where

$$\rho = \left(1 + \frac{\sigma^2}{n\tau_s^2}\right)^{-1/2}, \quad a = \frac{\theta_0 - \mu_s}{\tau_s}, \quad b = \rho\frac{k^{\pi^{(f)}}(n) - \mu_s}{\tau_s},$$

where $k^{\pi^{(f)}}(n)$ is as in (7) in Appendix A. Note that $\rho$ is always non-negative. The joint bivariate distribution comes from the joint probability distribution of $\bar{X}_n$ and $\theta$ as implied by the modelling of the likelihood and the prior. The quantity $a$ depends on the sampling prior alone while $b$ depends on the sampling prior, the fitting prior and the sample size $n$. The correlation between $\theta$ and $\bar{X}_n$ is $\rho$ which also depends on $n$.

In order to fix ideas, we provide a particular contour plot of the joint distribution of $U$ and $V$ in Figure 1. The two regions: (1) $U > a, V < b$ and (2) $U < a, V > b$ have been shaded. These two regions intersect at the point $(a, b)$. The location of the point $(a, b)$ and the shape of the contours of the bivariate normal distribution will change depending on the values of the sample size, $n$ and the prior parameters. Note, however, that the correlation will always be non-negative. The probabilities of these two regions under the bivariate normal distribution must be controlled to bound the risk function. How will it be possible to make the two probabilities very small? Unfortunately, there is no simple answer to this as the probabilities will depend on the actual prior parameters used and the sample size $n$ through $a$ and $b$. However, we provide the following theoretical and numerical results.

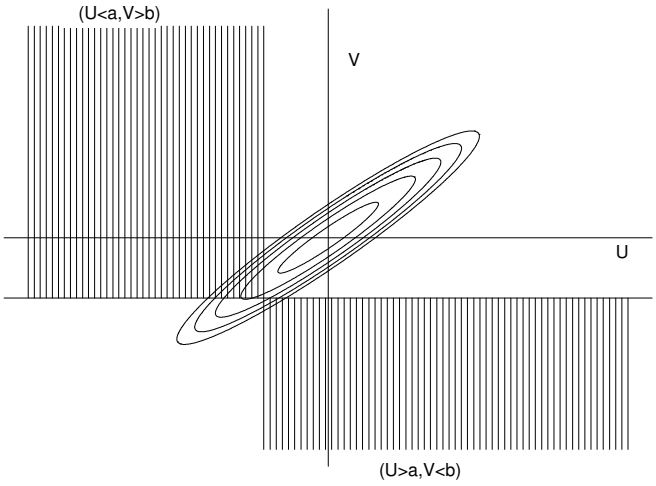


**Fig. 1.** A particular contour plot.

The two probabilities will be small (even for small $n$) if $a$ and $b$ are of the same sign, and both $|a|$ and $|b|$ are large. This happens when the point $(a, b)$ is far away from the origin in either direction along the major axis of the elliptical contours. When $a$ and $b$ are of opposite sign and at least one of $|a|$ and $|b|$ is large then one of the probabilities will be

zero and the other will be large for small values of $n$. Both the probabilities will be large for small $n$ if the point $(a, b)$ falls inside the high probability region of the contours. To reduce the high probabilities in the last two cases a large value of $n$ will be required. The large value of $n$ will make the value of $\rho$ close to 1 and as a result the contours will shrink to the major axis and both the probabilities will approach zero.

Suppose that $\tau_f^2$ is large corresponding to a non-informative fitting prior. Straightforward calculation yields that

$$b = \rho \left( a - \frac{q\,\sigma}{\tau_s \sqrt{n}} \right). \tag{4}$$

With a further assumption that $L_0 = L_1$, (equivalently $\eta = 1/2$) we have $q = 0$; now $b$ will be a positive multiple of $a$. Thus a large value of $|a|$ will yield a large value of $|b|$ of the same sign even for small values of $n$. As a result, even a very small sample size will be sufficient to make the two probabilities small. The quantity $a$ will be large if the mean of the sampling prior $\mu_s$ is quite far away from $\theta_0$ in units of $\tau_s$, the standard deviation of the sampling prior. Thus a smaller sample size can be expected if the prior mean is quite far away from the boundary value $\theta_0$ in either direction in units of $\tau_s$ when $L_0 = L_1$ and the variance of the fitting prior is large.

If we assume that both the sampling prior and the fitting prior are non-informative (in the sense that both $\tau_s^2$ and $\tau_f^2$ are large) then $b$ will be approximately equal to $a$ and as a result the sampling prior alone may dictate the sample size. That is, a smaller sample size can be expected if the prior mean is quite far away from the boundary value $\theta_0$ in either direction in units of $\tau_s$. Note that this conclusion does not require the equality assumption of the losses made in the preceding paragraph, since from (4) we have $b \to a$ as $\tau_s^2 \to \infty$ even when $q \neq 0$.

The two probabilities will be moderately large for small values of $n$ if the point $(a, b)$ is near the origin. The origin is the worst position of the point $(a, b)$ for making the probabilities of the two regions small since each of the two regions will intersect heavily with high probability areas of the bivariate normal distribution. Thus the $a = 0$ case for which the mean of the sampling prior is equal to $\theta_0$ will require a larger sample size than the $a \neq 0$ cases. The actual sample size, however, will depend on the magnitude of the quantity $b$ and the tightness of the upper bound on the risk function. An S-Plus routine to calculate the sample sizes is available from the corresponding author upon request.

We make several standard assumptions to illustrate the sample sizes in typical practical situations such as the one in Section 3.2 below.

A1: Without loss of generality we assume that $\theta_0 = 0$, thus the null hypothesis is $H_0 : \theta \leq 0$. Denoting $\theta$ to be the mean difference between the two rival treatments or procedures, this may mean that the new procedure or drug is not better than the current one. Note that the alternative hypothesis is $H_1 : \theta > 0$.

A2: We suppose that $L_0 = L_1$, i.e. $\eta = 1/2$ and $q = 0$. This implies that the loss function is symmetric.

A3: We assume that $\tau_f^2 = \sigma^2/n_f$ and $\tau_s^2 = \sigma^2/n_s$, so that $n_f$ and $n_s$ are the equivalent sample sizes implied by the fitting and sampling prior distributions, respectively.

The above choices lead to the following values of $\rho$, $a$, and $b$:

$$\rho = \sqrt{\frac{n}{n + n_s}},\ a = -\frac{\mu_s \sqrt{n_s}}{\sigma},\ b = -\rho \frac{\sqrt{n_s}}{\sigma} \left( \mu_s + \mu_f \frac{n_f}{n} \right). \tag{5}$$

Assuming A1, A2 and A3 we have the following results:

1. Suppose that $\mu_f = \mu_s = 0$ corresponding to the assumption that both priors are neither enthusiastic nor pessimistic about the two procedures since by assumption A1, $\theta_0 = 0$. This implies $a = b = 0$ and as a result the sample size will be solely determined by $n_s$, the prior sample size for the sampling prior. The exact number of samples will be determined by the rate at which $\rho$ approaches one, or equivalently $\frac{n_s}{n}$ approaches zero.

2. Suppose that the sampling and the fitting priors are the same; that is, $\mu_s = \mu_f = \mu_0$ and $n_f = n_s = n_0$. Then we have

$$\rho = \sqrt{\frac{n}{n+n_0}},\ a = -\frac{\mu_0\sqrt{n_0}}{\sigma},\ b = \frac{a}{\rho}.$$

   (a) Let $n_0 \to 0$ corresponding to limiting non-informative priors. Then both $a$ and $b$ approach zero and $\rho$ approaches one. A small sample size is required in this case, since the prior probability of $\theta$ being close to $\theta_0$ is very small and a small sample will indicate the actual location of $\theta$.
   (b) Let $n_0 \to \infty$ corresponding to a set of very informative priors, then $\rho$ approaches zero but both $|a|$ and $|b|$ will approach $\infty$ when $\mu_0 \neq 0$ and a small sample size is required as expected. If, however, $\mu_0 = 0$ then $a = b = 0$ and a very large sample size is required to guarantee that $\frac{n_0}{n}$ goes to zero, to have $\rho \to 1$. (Recall that the origin is the worst position of $(a,\ b)$ for SSD.) This also brings out a surprising finding that the optimal data sample size $n$ has to dominate the prior sample size $n_0$

From the above discussion it is clear that either a very small or a very large sample size is required for limiting prior distributions. In the following sub-section we consider a practical example and illustrate the sample sizes in more realistic situations where these theoretical results are re-confirmed. We also conduct experiments in the case where the fitting and sampling priors are different.

### *3.2. A clinical trial example*

Fayers *et al.* (2000) discuss the SSD problem for a trial for surgery for gastric cancer where a radical surgery (new treatment) is compared to conventional surgery (standard treatment). The log hazard ratio of death is the outcome of the trial and it follows an approximate normal distribution with mean $\theta$ and standard deviation, $\sigma = 2$, see Spiegelhalter *et al.* (2004, page 198) for justification of this assumption. The values of $\theta > 0$ favour the new treatment. In the classical setup the SSD problem is to determine $n$ such that the test of

$$H_0 : \theta = 0,\ \text{against}\ \ H_1 : \theta = \theta_a,$$

where $\theta_a$ is a fixed value specified as the alternative, at 5% significance level achieves 90% power. Fayers *et al.* (2000) discuss many different values of $\theta_a$ corresponding to some specific possible outcomes of the trials. By choosing $\theta_a = 0.29, 0.39$, and $0.56$, the approximate optimal sample sizes are 500, 276 and 134, respectively.

For the Bayesian problem, our hypotheses are of the form: $H_0 : \theta \leq 0$ and $H_1 : \theta > 0$, whereby the new surgery will be selected if the mean log hazard ratio is positive. Here the

magnitude of the expected treatment effect will not dictate the sample size unlike that in the classical setup where the alternative hypothesis plays a crucial role. We also assume that the losses are equal on the ground that an erroneous decision on either direction will incur the same amount of loss. This assumption is adopted in the absence of any information on the contrary.

Fayers *et al.* (2000) report prior opinions of 26 surgeons experienced in gastric surgery. By fitting a normal distribution on an appropriate transformed scale Spiegelhalter *et al.* (2004) conclude that the surgeons' opinion can be summarised by the $N(0.12, 0.19^2)$ prior distribution for $\theta$ which is an enthusiastic prior for the new treatment (radical surgery). This corresponds to $n_s = 111$ approximately since $\tau_s^2 = \sigma^2/n_s$. The power and level of significance requirements justify the value 0.15 for $M(\eta)$ in our implementation. The sample size we obtain using the proposed Bayesian method is 287 which is close to the sample size of 276 obtained using the classical method when $\theta_a = 0.39$. This, in our opinion, is a mere coincidence since apart from the same total error rate ($M(\eta) = 0.15$) the two procedures have little in common.

A single reported sample size is not very informative on its own and its sensitivity with respect to many different assumptions should be investigated. In a practical situation this sensitivity needs to be explored and matched with the practical information available to decide the sample size.

We first consider the case where the sampling and fitting prior distributions are the same, and we report the resulting sample sizes in Table 1. As expected the largest sample size is required for the $\mu_s = \mu_f = 0$ case (middle column in the table). This is expected because $a = b = 0$ in this case and recall that the origin is the worst position of $a$ and $b$ for SSD according to the discussion for Figure 1. Table 1 also reveals that sample size decreases as the prior mean moves away from $0(= \theta_0)$ in either direction, but the rate of decrease depends non-linearly on the assumed prior sample size. The sample size decreases as the prior sample size, $n_s = n_f$, decreases, i.e. a larger sample size is required for a tighter prior distribution.

It is clear from Table 1 that the sample size will be largest when both the prior means are equal to $\theta_0$, the null value of the mean (which is assumed to be zero here). This choice ($\mu_s = \mu_f = \theta_0$) has the potential to become a default case in many analyses since this corresponds to a prior distribution with mean which is neither enthusiastic nor pessimistic. Henceforth, we only investigate this case. Suppose that the sampling and fitting prior distributions have different variances (i.e. are based on different equivalent prior sample sizes, $n_s$ and $n_f$). In this case $n_f$ will not affect the optimum sample size since $b$ is free of $n_f$ and $n_f$ enters into the sample size calculation only through $b$, see equation (5) and note that $\mu_s = \mu_f = \theta_0 = 0$. Now the optimum sample size will depend on the value of $n_s$ and $M(\eta)$. The middle column (corresponding to $\mu_s = \mu_f = 0$) of Table 1 provides the numerical results. As before we continue to see that the sample sizes decrease with $n_s$. In conclusion, we recommend that a sensitivity study, like the one conducted here, should be undertaken before reaching a decision in any practical situation.

## 4. Application in financial audit

In auditing the final accounts about which a decision will be made comprise a set of sub-accounts such as, income (possibly by category) and expenditure on specific functions, for example payroll, or on products that are particular to the audited entity. Different sub-

**Table 1.** Optimum sample size for different values of prior mean (different columns) and prior sample sizes (different rows) for the clinical trial example when the fitting and sampling priors are same. Here $\theta_0 = 0$, $\eta = 1/2$ and $\sigma = 2$.

| | | | | | | $\mu_f = \mu_s$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | –0.60 | –0.50 | -0.40 | –0.30 | –0.20 | –0.10 | 0.00 | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 |
| $n_s = n_f$ | | | | | | $M(\eta) = 0.25$ | | | | | | | |
| 200 | 2 | 2 | 2 | 2 | 2 | 2 | 201 | 2 | 2 | 2 | 2 | 2 | 2 |
| 111 | 2 | 2 | 2 | 2 | 2 | 53 | 111 | 53 | 2 | 2 | 2 | 2 | 2 |
| 100 | 2 | 2 | 2 | 2 | 2 | 53 | 101 | 53 | 2 | 2 | 2 | 2 | 2 |
| 50 | 2 | 2 | 2 | 2 | 2 | 37 | 51 | 37 | 2 | 2 | 2 | 2 | 2 |
| 25 | 2 | 2 | 2 | 2 | 13 | 21 | 25 | 21 | 13 | 2 | 2 | 2 | 2 |
| 16 | 2 | 2 | 2 | 5 | 11 | 15 | 17 | 15 | 11 | 5 | 2 | 2 | 2 |
| 4 | 2 | 3 | 3 | 3 | 3 | 3 | 5 | 3 | 3 | 3 | 3 | 3 | 2 |
| 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | | | | | | $M(\eta) = 0.15$ | | | | | | | |
| 200 | 2 | 2 | 2 | 2 | 2 | 379 | 771 | 379 | 2 | 2 | 2 | 2 | 2 |
| 111 | 2 | 2 | 2 | 2 | 2 | 295 | 427 | 295 | 2 | 2 | 2 | 2 | 2 |
| 100 | 2 | 2 | 2 | 2 | 49 | 277 | 385 | 277 | 49 | 2 | 2 | 2 | 2 |
| 50 | 2 | 2 | 2 | 2 | 94 | 163 | 193 | 163 | 94 | 2 | 2 | 2 | 2 |
| 25 | 2 | 2 | 13 | 43 | 69 | 89 | 97 | 89 | 69 | 43 | 13 | 2 | 2 |
| 16 | 2 | 7 | 23 | 37 | 51 | 59 | 61 | 59 | 51 | 37 | 23 | 7 | 2 |
| 4 | 9 | 11 | 13 | 13 | 15 | 15 | 15 | 15 | 15 | 13 | 13 | 11 | 9 |
| 1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | | | | | | $M(\eta) = 0.10$ | | | | | | | |
| 200 | 2 | 2 | 2 | 2 | 2 | 1063 | 1895 | 1063 | 2 | 2 | 2 | 2 | 2 |
| 111 | 2 | 2 | 2 | 2 | 249 | 769 | 1051 | 769 | 249 | 2 | 2 | 2 | 2 |
| 100 | 2 | 2 | 2 | 2 | 269 | 715 | 947 | 715 | 269 | 2 | 2 | 2 | 2 |
| 50 | 2 | 2 | 2 | 109 | 265 | 411 | 473 | 411 | 265 | 109 | 2 | 2 | 2 |
| 25 | 2 | 17 | 67 | 123 | 179 | 221 | 237 | 221 | 179 | 123 | 67 | 17 | 2 |
| 16 | 17 | 43 | 71 | 101 | 127 | 145 | 151 | 145 | 127 | 101 | 71 | 43 | 17 |
| 4 | 25 | 29 | 31 | 35 | 37 | 37 | 37 | 37 | 37 | 35 | 31 | 29 | 25 |
| 1 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |

accounts have different accounting processes, and hence different types of error, and so the audit can be broken down into separate audits for each sub-account. If any sub-account is in serious error then the final audit conclusion will identify this and qualify this section of the accounts. Statistically the audit is stratified and inferences are made within strata as well as overall. Auditors use a concept called material error to define the value of monetary error that would lead them to qualify an account. We assume that the auditor has set the value of material error within each sub-account; typically this will be a percentage of the total money value of the sub-account, say 1% or 2%. Samples will be drawn from within strata and so we concentrate on SSD within each sub-account separately. In the rest of the paper the term account will refer to the sub-account being audited.

### 4.1. A mixture model

In financial audits the recorded value of a transaction is often called the *book value* which can be matched to a true value called the *audit value*. The error in a transaction is defined as the difference, $X_i' = B_i - A_i$, between its book value, $B_i$, and audit value, $A_i$. Often only overstatement errors can occur in which case we have $0 < A_i < B_i$ for all $i = 1, \ldots, n$. Following Cox and Snell (1979) we model the proportional errors, called the *taints*, $X_i = X_i'/B_i$.

Assume that $X_i$ is non-zero with probability $\psi$ and let there be $m$ items which result in positive errors. Denote these $m$ positive values of $X$ by $Z_1, Z_2, \ldots, Z_m$. Further, we assume that the random sample $Z_1, Z_2, \ldots, Z_m$ follows the exponential distribution with mean $\mu, 0 < \mu < 1$. Now the parameter of interest is given by $\theta = \psi\mu$, the proportion of error per money unit. The total error is $T_B\theta$ where $T_B = \sum B_i$ is the known total book value of the account.

As in Cox and Snell (1979) we assume that a-priori $\psi$ follows the gamma distribution with mean $\psi_0$, $G(a, a/\psi_0)$, and $\mu$ follows the inverse-gamma distribution with mean $\mu_0$, $IG(b, (b-1)\mu_0)$, independently for suitable values of $a$, $b$, $\psi_0$ and $\mu_0$. These prior distributions are adopted because they are conjugate, and as is well known a simpler analysis ensues under conjugate prior distributions. This simplification can also be justified by the fact that any SSD problem must involve a large number of assumptions and approximations. The joint prior density of $\psi$ and $\mu$ is given by:

$$\pi(\psi, \mu) = \left(\frac{a}{\psi_0}\right)^a \frac{1}{\Gamma(a)} \psi^{a-1} e^{-a\psi/\psi_0} \; \frac{\{(b-1)\mu_0\}^b}{\Gamma(b)} \frac{1}{\mu^{b+1}} e^{-(b-1)\mu_0/\mu}, \; \psi > 0, \mu > 0. \qquad (6)$$

After some calculation, we see that the induced prior distribution of the parameter of interest $\theta$, $\pi(\theta)$, is given by

$$\pi(\theta) = c\{\pi(\theta)\} \; F_{2a,2b} \text{ where } c\{\pi(\theta)\} = \left\{\frac{(b-1)\psi_0\mu_0}{b}\right\},$$

and $F_{\nu_1,\nu_2}$ is the standard $F$ random variable with $(\nu_1, \nu_2)$ degrees of freedom.

Note that the prior mean of $\theta = \psi\mu$ is given by the product $\psi_0\mu_0$; the other hyper-parameters $a$ and $b$ cancel out in the mean. However, the variance of $\theta$ depends on all the hyper-parameters and we shall return to their choices later.

The likelihood is obtained by arguing that $m|n, \psi$ follows the Poisson distribution with parameter $n\psi$ and given $m > 0$, $Z_1, \ldots, Z_m$ are i.i.d. exponential random variables with

mean $\mu$. The resulting likelihood is given by:

$$L(\psi, \mu; n, m, \mathbf{z}) \propto e^{-n\psi}(n\psi)^m \frac{1}{\mu^m} e^{-\frac{1}{\mu}\sum_{i=1}^m z_i}.$$

The joint posterior distribution of $\psi$ and $\mu$ is proportional to $L(\psi, \mu; n, m, \mathbf{z}) \times \pi(\psi, \mu)$, and is given by

$$\pi(\psi, \mu | n, m, \mathbf{z}) \propto e^{-n\psi}(n\psi)^m \frac{1}{\mu^m} e^{-\frac{1}{\mu}\sum_{i=1}^m z_i} \psi^{a-1} e^{-a\psi/\psi_0} \frac{1}{\mu^{b+1}} e^{-(b-1)\mu_0/\mu},$$

for $\psi > 0$ and $\mu > 0$. If $m = 0$ then we simply drop the terms involving $m$ from the above expression to obtain the posterior distribution.

After some integration, we see that the posterior distribution of the quantity $\theta = \psi\mu$ is given by

$$\pi(\theta | \mathbf{x}^{(n)}) = c\left\{\pi(\theta | \mathbf{x}^{(n)})\right\} F_{2(m+a),2(m+b)},$$

where

$$c\left\{\pi(\theta | \mathbf{x}^{(n)}\right\} = \left\{\frac{m\bar{z}_m + (b-1)\mu_0}{n + a/\psi_0}\right\}\left(\frac{m+a}{m+b}\right).$$

Note that if $m = 0$ then the posterior distribution is given by

$$\pi(\theta | \mathbf{x}^{(n)}) = c\left\{\pi(\theta | \mathbf{x}^{(n)})\right\} F_{2a,2b}, \text{ where } c\left\{\pi(\theta | \mathbf{x}^{(n)})\right\} = \frac{a}{b}\left\{\frac{(b-1)\mu_0}{n + a/\psi_0}\right\}.$$

Further, when $n = 0$ it is easy to see that the prior and posterior distributions of $\theta$ coincide, as expected. The technical details for estimating the sample sizes are given in Appendix B.

### *4.2. Numerical results*

The prior mean and variance of $\theta$ are given by:

$$\text{mean} = \psi_0\mu_0, \quad \text{variance} = \frac{a+b-1}{a(b-2)}(\psi_0\mu_0)^2.$$

We express our prior parameter values in units of the auditor's material error, $\theta_0$, as follows. We assume $\psi_0 = 0.01$ and obtain values of $\mu_0$ using the relation $\psi_0\mu_0 = k_1\theta_0$ for different values of $k_1$. We now set the prior standard deviation at $k_2$ times $\theta_0$, i.e.

$$\left(\frac{a+b-1}{a(b-2)}\right)^{1/2} \psi_0\mu_0 = k_2\theta_0.$$

This provides only one constraint for two undetermined parameters $a$ and $b$, so many different strategies can be adopted. In order to ensure positivity of both $a$ and $b$ we require that

$$b > 2 + \frac{k_1^2}{k_2^2}.$$

We let

$$b = 2 + \frac{k_1^2}{k_2^2} + b_0 \text{ and } a = \frac{b-1}{k_2^2(b-2)/k_1^2 - 1},$$

where $b_0$ is a non-negative parameter. A small value of $b_0$ makes the prior distribution very spiky and as a result the sample sizes become very large. That is why we illustrate with a moderate value of $b_0 = 10$, although other values can be adopted.

In our illustration, we assume that $\psi_0^{(s)} = \psi_0^{(f)} = 0.01$ to reduce the number of parameters to be given as input for the method. The remaining parameters in the prior distributions are obtained by specifying particular values for $k_1$ and $k_2$. Note that we shall have four parameters $k_1^{(f)}, k_2^{(f)}, k_1^{(s)}$ and $k_2^{(s)}$ for the fitting and sampling priors.

- Suppose that the sampling and the fitting priors are the same. In this case we have $a_s = a_f$ and $b_s = b_f$. Note that these parameters are obtained by first assuming a particular value for each of $k_1^{(s)} = k_1^{(f)} = k_1$ and $k_2^{(s)} = k_2^{(f)} = k_2$. The optimal sample sizes are reported in Table 2. Here the sample sizes are not symmetric around the $k_1 = 1$ column due to skewness of the mixture distribution. The sample sizes decrease when the prior variance increases as in the normal case. Also note that there are some optimal sample sizes which are $\infty$. These are due to the corresponding very small prior variances assumed. The implied prior distribution for each of these cases resembles a spike (centered very close to $\theta_0$) and huge number of samples are required to discriminate between the two hypotheses. In practical auditing terms these infinite sample sizes will require a complete audit.

- In Table 3 we assume that $k_1^{(s)} = k_1^{(f)} = 1$, but we specify different values of $k_2^{(s)}$ and $k_2^{(f)}$ for the sampling and fitting prior. As in the previous clinical trial example the optimum sample sizes are not affected by the fitting prior distribution; the small variation between the columns is due to sampling fluctuations in the simulation. Also as seen previously higher sample sizes are needed for tighter sampling prior distributions (see the variations between the rows of the table).

- Now we suppose that there is a mismatch between the means of the fitting and sampling prior distributions. To illustrate we assume that $k_1^{(s)} = 0.5$ and $k_1^{(f)} = 1$. We report the optimum sample sizes in Table 4 for different values of $k_2^{(s)}$ and $k_2^{(f)}$ but only for $M(\eta) = 0.1$. For the other two values of $M(\eta)$ the sample sizes were trivially small. The optimum sample size decreases when $k_2^{(s)}$ increases and is not affected a great deal by the variance of the fitting prior when the variance of the sampling prior is moderately large.

## 5. Discussion

There are so many uncertainties in SSD that approximate methods have to be employed within any theoretical framework. In this article we have explored some of the implication of this within a full Bayesian framework for SSD. Our approach is general and can be used for many problems in statistical decision making. We have found that typical non-informative prior distributions lead to very small sample sizes. On the other hand, a very informative prior distribution also leads to a very small sample size when the prior mean is 'far away' from the hypothesized value of the parameter. The sample sizes are the largest when the prior distribution concentrates very strongly at the hypothesized value of the parameter. These results have been shown both theoretically and numerically. An S-Plus routine is available from the corresponding author upon request.

**Table 2.** Optimum sample size for different values of $k_1$ and $k_2$ for the mixture example when the fitting and sampling prior are same. Here $\theta_0 = 0.01$, $\eta = 1/2$, $\psi_0^{(s)} = \psi_0^{(f)} = 0.01$ and $\mu_0^{(s)} = \mu_0^{(f)}$.

| | | | | $k_1$ | | | |
|---|---|---|---|---|---|---|---|
| | 0.25 | 0.5 | 0.75 | 1.0 | 1.25 | 1.5 | 1.75 |
| $k_2$ | | | | $M(\eta) = 0.25$ | | | |
| 0.5 | 2 | 2 | 3 | 573 | 903 | 2 | 2 |
| 1.0 | 2 | 2 | 2 | 75 | 215 | 264 | 2 |
| 1.5 | 2 | 2 | 2 | 15 | 49 | 111 | 146 |
| 2.0 | 2 | 2 | 2 | 3 | 22 | 53 | 83 |
| 2.5 | 2 | 2 | 2 | 2 | 8 | 28 | 54 |
| | | | | $M(\eta) = 0.15$ | | | |
| 0.5 | 2 | 2 | 426 | 2387 | 3646 | 2 | 2 |
| 1.0 | 2 | 2 | 86 | 398 | 822 | 1257 | 1008 |
| 1.5 | 2 | 2 | 20 | 94 | 293 | 555 | 648 |
| 2.0 | 2 | 2 | 6 | 31 | 123 | 195 | 416 |
| 2.5 | 2 | 2 | 2 | 14 | 56 | 137 | 219 |
| | | | | $M(\eta) = 0.10$ | | | |
| 0.5 | 2 | 65 | 1500 | $\infty$ | $\infty$ | 3341 | 2 |
| 1.0 | 2 | 37 | 275 | 1011 | 2285 | 3153 | 2809 |
| 1.5 | 2 | 6 | 80 | 307 | 746 | 1405 | 1871 |
| 2.0 | 2 | 2 | 27 | 103 | 318 | 701 | 1017 |
| 2.5 | 2 | 2 | 12 | 43 | 131 | 317 | 589 |

**Table 3.** Optimum sample size for different values of $k_2^{(s)}$ and $k_2^{(f)}$ for the mixture example. Here $\theta_0 = 0.01$; $k_1^{(s)} = k_1^{(f)} = 1$, $\psi_0^{(s)} = \psi_0^{(f)} = 0.01$, $\eta = 1/2$.

| | | | $k_2^{(f)}$ | | |
|---|---|---|---|---|---|
| | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 |
| $k_2^{(s)}$ | | | $M(\eta) = 0.25$ | | |
| 0.5 | 497 | 595 | 592 | 556 | 490 |
| 1.0 | 72 | 79 | 76 | 78 | 51 |
| 1.5 | 15 | 16 | 15 | 14 | 14 |
| 2.0 | 3 | 3 | 3 | 4 | 5 |
| 2.5 | 2 | 2 | 2 | 2 | 2 |
| | | | $M(\eta) = 0.15$ | | |
| 0.5 | 2260 | 2334 | 2230 | 2254 | 2256 |
| 1.0 | 420 | 368 | 328 | 357 | 389 |
| 1.5 | 131 | 107 | 106 | 121 | 105 |
| 2.0 | 46 | 34 | 37 | 29 | 37 |
| 2.5 | 24 | 13 | 15 | 18 | 14 |
| | | | $M(\eta) = 0.10$ | | |
| 0.5 | $\infty$ | $\infty$ | $\infty$ | $\infty$ | $\infty$ |
| 1.0 | 1086 | 1032 | 1058 | 1126 | 1039 |
| 1.5 | 337 | 334 | 304 | 301 | 325 |
| 2.0 | 128 | 142 | 99 | 129 | 114 |
| 2.5 | 75 | 66 | 64 | 50 | 48 |

**Table 4.** Optimum sample size for different values of $k_2^{(s)}$ and $k_2^{(f)}$ for the mixture example. Here $\theta_0 = 0.01$; $k_1^{(s)} = 0.5, k_1^{(f)} = 1$, $\psi_0^{(s)} = \psi_0^{(f)} = 0.01$, $\eta = 1/2$.

| | $k_2^{(f)}$ | | | | |
|---|---|---|---|---|---|
| | 0.5 | 1.0 | 1.5 | 2.0 | 2.5 |
| $k_2^{(s)}$ | $M(\eta) = 0.10$ | | | | |
| 0.5 | 499 | 984 | 1003 | 991 | 971 |
| 1.0 | 37 | 58 | 63 | 65 | 63 |
| 1.5 | 4 | 6 | 5 | 5 | 8 |
| 2.0 | 2 | 3 | 2 | 2 | 3 |
| 2.5 | 2 | 2 | 2 | 2 | 2 |

The results for the normal distribution apply to a wide range of applications, including the clinical trial example that we have chosen. We feel that the Bayesian framework can incorporate practitioners prior knowledge regarding the hypotheses and potential losses far more naturally than those required in a frequentist framework.

A key result in the auditing context is that if the prior mean is far away from the boundary value, $\theta_0$ (or the per item material error), then the required sample size is very small which confirms the auditors' views about the value of sampling. In this case a minimum sample size should be set to satisfy auditing standards and to guarantee some level of quality assurance due to sampling. If the prior mean is very close to the material error then, as expected, a large sample size is required. This sample size gets even larger for the tighter prior distributions. Also when the upper bound on the two error probabilities, $M(\eta)$, is small the sample sizes become very large.

It should be noted that the substantive testing of items tests only the accuracy of the totals generated by the system as specified at the first stage of the audit. This is not a procedure designed to discover large faults in the design of the system that may have led to recent accounting scandals such as those at Enron and Parmalat. Discovering these system faults is the responsibility of senior auditors at the system review stage.

The optimal sample sizes in the two examples have been found under two different parametric assumptions on the error distribution, but the key conclusions remained the same across the two models. The sample sizes are model dependent if the prior mean of $\theta$ is close to the hypothesized value $\theta_0$. However, if the prior mean is very 'far away' from $\theta_0$, which is often the case, both the models give very small sample sizes.

Lastly, we feel that a clear distinction should be made between the sampling and fitting prior distributions. The sampling prior distribution relates to the data generating mechanism while the fitting prior drives the inference through the posterior distribution. Intuition suggests that a non-informative fitting prior distribution should not influence the sample size and we have demonstrated this here. The sampling prior distribution captures the practitioners' usually strong prior belief while the fitting prior distribution is a statistician's device to implement the analysis.

## Appendix A: Calculations for the normal likelihoods in Section 3.1.

We recall that $X|\theta \sim N(\theta, \sigma^2)$ where $\sigma^2$ is known and assume $\pi^{(f)}(\theta) = N(\mu_f, \tau_f^2)$ and $\pi^{(s)}(\theta) = N(\mu_s, \tau_s^2)$. The posterior distribution of $\theta$ is normal with mean

$$E(\theta|\bar{x}_n) = \lambda_f^2 \left( \frac{n\bar{x}_n}{\sigma^2} + \frac{\mu_f}{\tau_f^2} \right), \text{ and } \mathrm{var}(\theta|\bar{x}_n) = \lambda_f^2$$

where $\lambda_f^2 = 1/\left(\frac{n}{\sigma^2} + \frac{1}{\tau_f^2}\right)$. We now derive $k^{\pi^{(f)}}(n)$. The Bayes rule chooses action $a_0$ if

$$\begin{array}{lccl} & L_0 \int_{\theta_0}^{\infty} \pi^{(f)}(\theta|\bar{x}_n)d\theta & < & L_1 \int_{-\infty}^{\theta_0} \pi^{(f)}(\theta|\bar{x}_n)d\theta \\ i.e. & L_0(1-p) & < & L_1 p, \text{ say,} \\ \Longrightarrow & p & > & \frac{L_0}{L_0+L_1} \equiv \eta, \end{array}$$

where

$$p = \int_{-\infty}^{\theta_0} \pi^{(f)}(\theta|\bar{x})d\theta = \Phi\left( \frac{\theta_0 - \lambda_f^2(\frac{n\bar{x}_n}{\sigma^2} + \frac{\mu_f}{\tau_f^2})}{\lambda_f} \right),$$

and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Let $\Phi^{-1}$ denote the inverse of $\Phi$ and $q = \Phi^{-1}(\eta)$. Now it is clear that $p > \eta$ if

$$\bar{x}_n < k^{\pi^{(f)}}(n) = \frac{\sigma^2}{n} \left\{ \frac{\theta_0 - q\lambda_f}{\lambda_f^2} - \frac{\mu_f}{\tau_f^2} \right\}. \tag{7}$$

We now have

$$P(\bar{X}_n < k^{\pi^{(f)}}(n)|\theta) = \Phi\left( \frac{k^{\pi^{(f)}}(n) - \theta}{\sigma/\sqrt{n}} \right).$$

Let $\phi(\cdot)$ be the density function of the standard normal random variable. The following calculations reduce the risk function to an analytic form. The risk is given by

$$\begin{array}{rcl} r(\pi^{(s)}, \delta_n^{\pi^{(f)}}) & = & L_0 \int_{\theta_0}^{\infty} \Phi\left( \frac{k^{\pi^{(f)}}(n)-\theta}{\sigma/\sqrt{n}} \right) \frac{1}{\tau_s\sqrt{2\pi}} e^{-\frac{1}{2\tau_s^2}(\theta-\mu_s)^2} d\theta \\ & & + L_1 \int_{-\infty}^{\theta_0} \left\{ 1 - \Phi\left( \frac{k^{\pi^{(f)}}(n)-\theta}{\sigma/\sqrt{n}} \right) \right\} \frac{1}{\tau_s\sqrt{2\pi}} e^{-\frac{1}{2\tau_s^2}(\theta-\mu_s)^2} d\theta, \\ & = & L_0 \int_{\frac{\theta_0-\mu_s}{\tau_s}}^{\infty} \Phi\left( \frac{k^{\pi^{(f)}}(n)-\mu_s-\tau_s u}{\sigma/\sqrt{n}} \right) \phi(u)du \\ & & + L_1 \int_{-\infty}^{\frac{\theta_0-\mu_s}{\tau_s}} \left\{ 1 - \Phi\left( \frac{k^{\pi^{(f)}}(n)-\mu_s-\tau_s u}{\sigma/\sqrt{n}} \right) \right\} \phi(u)du \\ & = & L_0 P(U^* < -a, V^* < b) + L_1 P\left(U^* < a, V^* < -b\right) \end{array}$$

where

$$a = \frac{\theta_0 - \mu_s}{\tau_s},\ b = \frac{d}{\sqrt{1+c^2}},\ c = -\frac{\sqrt{n}\tau_s}{\sigma},\ \text{and } d = \frac{\sqrt{n}\left(k^{\pi^{(f)}}(n) - \mu_s\right)}{\sigma}.$$

and $U^*$ and $V^*$ jointly follow the bivariate normal distribution with zero means, unit variances and correlation $\rho^* = \frac{c}{\sqrt{1+c^2}}$. We have used the following two identities:

$$\begin{aligned} \int_a^{\infty} \phi(z)\Phi(cz+d)dz &= P\left(U^* < -a, V^* < b\right) \\ \int_{-\infty}^{a} \phi(z)\left(1-\Phi(cz+d)\right)dz &= P\left(U^* < a, V^* < -b\right). \end{aligned} \tag{8}$$

These two results are proved similarly, the proof of the first identity (8) is given below.

We have

$$\int_a^{\infty} \phi(z)\Phi(cz+d)dz = \int_a^{\infty} \phi(z) \int_{-\infty}^{cz+d} \phi(y)dydz = \int_{-\infty}^{-a} \phi(z) \int_{-\infty}^{-cz+d} \phi(y)dydz.$$

Now we work with the right hand side as follows:

$$\begin{aligned} P(U^* < -a, V^* < \tfrac{d}{\sqrt{1+c^2}}) &= \int_{-\infty}^{-a} \int_{-\infty}^{\frac{d}{\sqrt{1+c^2}}} \frac{1}{2\pi\sqrt{1-\rho^{*2}}} e^{-\frac{1}{2(1-\rho^{*2})}\left(u^2-2\rho^* uv+v^2\right)} dudv \\ &= \int_{-\infty}^{-a} \int_{-\infty}^{\frac{1}{\sqrt{1-\rho^{*2}}}\left(\frac{d}{\sqrt{1+c^2}}-\rho^* z\right)} \phi(y)\phi(z)dydz, \\ &= \int_{-\infty}^{-a} \phi(z) \int_{-\infty}^{-cz+d} \phi(y)dydz. \end{aligned}$$

by using the transformation $z = u$, $y = \frac{1}{\sqrt{1-\rho^{*2}}}(v - \rho^* u)$, and then by substituting the value of $\rho^*$. This completes the proof.

By applying a further transformation we re-write the risk function as:

$$r(\pi^{(s)}, \delta_n^{\pi^{(f)}}) = L_0 P\left(U > a, V < b\right) + L_1 P\left(U < a, V > b\right), \tag{9}$$

where $U$ and $V$ jointly follow the bivariate normal distribution with zero means, unit variances and correlation

$$\rho = \left(1 + \frac{\sigma^2}{n\tau_s^2}\right)^{-1/2},$$

and

$$a = \frac{\theta_0 - \mu_s}{\tau_s}, \quad b = \rho\frac{k^{\pi^{(f)}}(n) - \mu_s}{\tau_s}.$$

Thus we have an analytic expression for the risk function which can be evaluated for different values of the sample size $n$ and the optimum can be found.

## Appendix B: Calculations for the mixture model in Section 4.1.

The Bayes rule chooses action $a_0$ if

$$\int_0^{\theta_0} \pi(\theta|\mathbf{x}^{(n)})d\theta > \frac{L_0}{L_0 + L_1} \equiv \eta,$$

as before. This holds if,

$$\frac{\theta_0}{c\left\{\pi(\theta|\mathbf{x}^{(n)})\right\}} \geq q(m, a, b, \eta), \tag{10}$$

where $q(m, a, b, \eta)$ satisfies

$$P\left\{F_{2(m+a),2(m+b)} < q(m, a, b, \eta)\right\} = \eta.$$

For the inequality (10), two cases arise depending on the value of $m$. If $m > 0$, then the Bayes rule chooses action $a_0$ if

$$\sum_{i=1}^{m} z_i < \theta_0 \frac{m+b}{m+a} \frac{n + a/\psi_0}{q(m, a, b, \eta)} - (b-1)\mu_0. \tag{11}$$

On the other hand, if $m = 0$ then the Bayes rule chooses action $a_0$ if

$$\theta_0 \frac{b}{a} \frac{n + a/\psi_0}{q(0, a, b, \eta)} > (b-1)\mu_0. \tag{12}$$

Consequently, depending on the value of $m$ the probability $P\left\{g\left(\mathbf{X}^{(n)}\right) < k(n)|\theta\right\}$ will have two different forms. When $m = 0$, the probability is 1 if (12) is satisfied and 0 otherwise. If, however, $m$ is non-zero then the probability is given by

$$P\left(Y < \frac{\theta_0}{\mu} \frac{m+b}{m+a} \frac{n + a/\psi_0}{q(m, a, b, \eta)} - (b-1)\frac{\mu_0}{\mu}\right)$$

where $Y$ follows the gamma distribution $G(m, 1)$. Note that this probability will be zero when the right hand side of (11) is negative.

We now introduce the fitting and the sampling priors for calculating the risk function (3). Assume that the forms of the fitting and sampling prior distributions are the same. Let $a_f, b_f, \psi_0^{(f)}, \mu_0^{(f)}$ be the parameters under the fitting prior and $a_s, b_s, \psi_0^{(s)}, \mu_0^{(s)}$ be the parameters under the sampling prior. Now the probabilities $P\left\{g\left(\mathbf{X}^{(n)}\right) < k(n)|\theta\right\}$ and $P\left\{g\left(\mathbf{X}^{(n)}\right) \geq k(n)|\theta\right\}$ are to be calculated using the parameter values $a_f, b_f, \psi_0^{(f)}, \mu_0^{(f)}$ for the fitting prior.

The risk function (3) is now calculated using Monte Carlo sampling from the sampling prior distribution as follows. We first simulate $\psi$ and $\mu$ from their sampling prior distributions which have hyper-parameters $a_s, b_s, \psi_0^{(s)}, \mu_0^{(s)}$. The product $\theta = \psi\mu$ is taken as a draw from the sampling prior distribution. Conditional on the draws from the prior distribution we simulate $m$ for a given sample size $n$ using the fact that $m|n, \psi$ follows the Poisson distribution with parameter $n\psi$.

The probability of choosing actions $a_0$ and $a_1$ are evaluated under the fitting prior distributions which have hyper-parameters $a_f, b_f, \psi_0^{(f)}, \mu_0^{(f)}$. That is, we set

$$P\left\{g\left(\mathbf{X}^{(n)}\right) < k^{\pi^{(f)}}(n)|\theta\right\} = \begin{cases} I\left(\theta_0 \frac{b_f}{a_f} \frac{n+a_f/\psi_0^{(f)}}{q(0,a_f,b_f,\eta)} > (b_f-1)\mu_0^{(f)}\right), & \text{if } m = 0 \\ G_m\left(\frac{\theta_0}{\mu} \frac{m+b_f}{m+a_f} \frac{n+a_f/\psi_0^{(f)}}{q(m,a_f,b_f,\eta)} - (b_f-1)\frac{\mu_0^{(f)}}{\mu}\right), & \text{otherwise,} \end{cases}$$

where $I(\cdot)$ denotes the indicator function. Subsequently the average risk over 2000 simulation replications produces accurate estimates of the risk $r\left(\pi^{(s)}, \delta_n^{\pi^{(f)}}\right)$.

## Acknowledgments

The authors would like to thank Prof Vic Barnett, John Haworth, Robin Swan and colleagues of the National Audit Office for many helpful discussions and suggestions.

## References

Adcock, C. J. (1997). Sample size determination: a review. *The Statistician 46*, 261–283.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Clarke, B. S. and A. Yuan (2002). A closed form expression for Bayesian sample sizes. Technical report, Department of Statistics, University of British Columbia.

Cox, D. R. and E. J. Snell (1979). On sampling and the estimation of rare errors. *Biometrika 66*, 125–132.

Fayers, P. M., A. Cushieri, J. Fielding, B. Uscinska, and L. S. Freedman (2000). Sample size calculation for clinical trials: the impact of clinician beliefs. *British Journal of Cancer 82*, 213–219.

Heiner, K. W. and O. Whitby (1980). Maximizing restitution for erroneous medical payments when auditing samples. *Interfaces 10*, 46–54.

Joseph, L., D. B. Wolfson, and R. du Berger (1995). Sample size determination for binomial proportions via highest posterior density intervals. *The Statistician 44*, 143–154.

Laws, D. J. and A. O'Hagan (2000). Bayesian inference for rare errors in populations with unequal unit sizes. *Applied Statistics 49*, 577–590.

Laws, D. J. and A. O'Hagan (2002). A hierarchical Bayesian model for rare errors. *The Statistician 51*, 431–450.

Lindley, D. V. (1997). The choice of sample size. *The Statistician 46*, 129–138.

Patterson, E. R. (1993). Strategic Sample-size choice in auditing. *Journal of Accounting Research 31*, 272–293.

Raiffa, H. and R. Schlaifer (2000). *Applied Statistical Decision Theory (Wiley Classics Library)*. London: John Wiley & Sons.

Shrivastava, R. P. and G. R. Shafer (1994). Integrating statistical and non-statistical audit evidence using belief functions - a case of variable sampling. *International journal of intelligent systems 9*, 519–539.

Smith, T. M. F. (1976). *Statistical Sampling for Accountants: Accountancy Age Books*. London: Haymarket Publishing Limited.

Smith, T. M. F. (1979). Statistical sampling in auditing: a statistician's view point. *The Statistician 28*, 267–280.

Spiegelhalter, D. J., K. R. Abrams, and J. P. Myles (2004). *Bayesian Approaches to Clinical Trials and Health-Care Eavaluation*. London: John Wiley & Sons, Ltd.

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

Spiegelhalter, D. J., L. S. Freedman, and M. K. B. Parmar (1994). Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society, Series A 157*, 357–416.

Walker, S. G. (2003). How many samples?: a Bayesian non-parametric approach. *The Statistician 52*, 475–482.

Wang, F. and A. E. Gelfand (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science 17*, 193–208.