

SAMPLE SIZE PROBLEMS IN ANOVA: BAYESIAN POINT OF
VIEW

by

Anirban DasGupta
Purdue University

and

Brani Vidakovic
Duke University

Technical Report #94-19C

Department of Statistics
Purdue University

1994

SAMPLE SIZE PROBLEMS IN ANOVA: BAYESIAN POINT OF VIEW¹

BY
ANIRBAN DASGUPTA
Purdue University
AND
BRANI VIDAKOVIC
Duke University

In this paper we discuss the sample size problem for balanced one way ANOVA under a posterior Bayesian formulation of the problem. Using the distribution theory of appropriate quadratic forms we derive explicit sample sizes for prespecified posterior precisions. Comparisons with classical sample sizes are made. Instead of extensive tables, a MATHEMATICA program for sample size calculation is given. The formulations given in this article form a foundational step towards Bayesian calculation of sample size, in general.

Key words and phrases: Sample size problem, ANOVA, Bayesian point of view.

AMS Subject Classification: 62F05, 62F15.

1 Introduction

The decision about the size of a sample is one of the main components in planning a statistical inference. Too small a sample may give unsatisfactory statistical results while an unnecessarily large sample wastes money and resources. The problem of determining sample sizes that give an assurance of a prespecified accuracy (in whatever relevant way accuracy is defined) has assumed the status of textbook material in classical statistics. Typically, for instance, one may ask for a power $(1 - \beta)$ of .95 at a type 1 error level of $\alpha = .05$ in standard hypotheses testing problems. We routinely teach our students in most elementary classes about such classical sample sizes. Several books and monographs testify to the value that practitioners assign to sample sizes as a preexperimental design component; see for instance Kestanbaum et al. (1970) and Odeh (1991).

Two main steps dominate in the procedure for determining the sample size:

- (i) A requirement in terms of desired *precision*
- (ii) A *mathematical connection* between the sample size n and the required precision in a form of an implicit or explicit equation must be determined. By solving this equation the minimal sample size is obtained.

¹Research supported by NSF Grant DMS-9307727 at Purdue University and NSF Grant DMS-9404151 at Duke University.

In the case of random effects ANOVA, the error of second kind, β , can be expressed in terms of a central F distribution, while in a fixed effects ANOVA, β is a quantile of a non-central F distribution. The power $1 - \beta$ is compared with the non-centrality parameter Φ that measures the extent to which H_0 is false. Since Φ depends on n , fixing the power against some common alternatives gives the required sample size. Charts for determining the sample size are given in Pearson and Hartley (1951).

It is somewhat curious that barring a recent outgrowth of interest and activity, determination of sample sizes correct according to a Bayesian formulation has been more or less a nonexistent topic in Bayesian research. First, one needs to understand that a new and careful formulation consistent with a Bayesian view is necessary. The Bayesian has little primary use for sample sizes that assure low type 1 and type 2 error probabilities. In the Bayesian's mind, the correct accuracy measure is a posterior accuracy: one accepts the hypothesis with a larger posterior probability. At the same time it is desirable that the rejected hypothesis has small posterior probability. That is the criterion that will lead to a sample size determination.

Let y be a vector of sufficient statistics for the problem. The posterior risk (under the '0-1' loss) is $\min\{Pr(H_0|y), Pr(H_1|y)\}$ and the condition

$$\min\{Pr(H_0|y), Pr(H_1|y)\} \leq \epsilon \quad (1)$$

imposed on the risk is a natural one. The rejected hypothesis should have small posterior probability. This, however, is rather subtle; sample size determination is a preexperimental exercise. What data will come our way is anybody's guess; posterior accuracies are by nature functions of the obtained data. One can and must therefore either seek an assurance of posterior accuracy for all possible data or at least all probable data: data that are likely to be seen once the prescribed sample size is used and data are obtained.

Condition (1), however, is usually impossible to satisfy for all possible data y . In such a case, a set T should be chosen such that

$$Pr(y \notin T) = Pr(y \in K) < \delta, \quad (2)$$

for some small δ , where $K = T^c$ and condition (1) is satisfied for all data y in the set T . The probability in (2) is expressed in terms of the marginal distribution of y .

It is important to understand that T is not a preferred set of samples; it is simply an appropriate set of samples for which a prespecified posterior accuracy is guaranteed and using such a set T is essential in the Bayesian formulation of the sample size problem because a guaranteed posterior accuracy for all samples is simply impossible. The results here thus give a preposterior guarantee of posterior accuracy with a large predictive probability. If by chance the obtained data happen to be one of the samples we were not preprotected against, the automatic accuracy is invalidated. Of course, that may also mean the modeling was wrong and careful rethinking about the model is necessary: why did such unanticipated data even arise? The exact sample size is found by solving a moderately messy equation treating the sample size n as a variable. The user uses his/her own parameters for the prior, and a computer code built for the purpose solves the relevant equation.

Several critical questions deserve thinking. Is it important to have flexibility in the functional form of the prior as well? In principle, the answer has to be 'yes'. Who knows

what prior is deemed appropriate in a given problem? But it is quite plainly a fact that the associated mathematics changes completely with a change in the functional form of the prior. Also, confronted with prescriptions of different sample sizes for different kinds of priors, it is very likely that the reaction will be one of resignation. It therefore appears to be natural that one writes a computer code for Bayesian sample sizes only for conjugate priors: the code will allow flexibility in the hyperparameters of the prior to suit the need of the individual user.

It is a natural curiosity and question of simple pragmatism to ask if adoption of the classical sample size will provide any kind of acceptable posterior accuracy. If in a series of problems, evidence accumulates that classical sample sizes are not really too bad as Bayesian sample sizes, it would be, realistically and pragmatically, time to stop solving complicated mathematics problems and writing involved computer codes for Bayesian sample sizes. As it turns out, quite typically, classical sample sizes do not seem to provide very good posterior accuracy; further elaboration and discussion of this issue is given in Section 4.

Section 2 describes the exact problem and the exact priors we use and identifies the set T . This is done by deriving a formula for the Bayes factor. Section 3 gives the exact equation to be solved for determination of the sample size by using distributional theory of appropriate quadratic forms. Section 4 gives some explicit examples, and comparison with classical sample sizes. We conclude with some brief general remarks. A MATHEMATICA code for general use is given in the Appendix.

Hutton and Owens (1993) suggest Bayesian designing of sample size through the lengths of credible sets. DasGupta and Mukhopadhyay (1994) address the problem of determining the sample size in Bayesian fashion which ensures robust posterior inference. Sample size problems from the Bayesian perspective are also discussed in Berger (1985) and Lee et al (1993). Of course, the Bayesian sample size problem is really a Bayes design problem; the literature on Bayes design is quite rich and includes Chaloner (1984), DasGupta and Studden (1991), Pilz (1991), among others.

This is a first formal attempt to give Bayesian sample sizes in one way ANOVA, one of the most used statistical methods. We give a clear theory, give explicit sample sizes and give a computer code for sample size calculations for general use. Tools such as Bayesian sample sizes may look esoteric to some right now, but will be even more valuable to practitioners as Bayesian statistics becomes more accepted. Applied Bayesians would naturally want a guideline regarding sample size in actual problems, and this article exactly gives them such a guideline. Bayesian sample sizes are also of value to classical statisticians; if Bayesian sample sizes for reasonable prior distributions severely conflict with sample sizes being used by a classical statistician, there is a reason for concern and reevaluation of the sample size being used in the problem.

Other models are currently under active investigation.

2 Mathematics of the problem

We consider a simple case of balanced one-way ANOVA where the variance of the observations, σ^2 , is given. Since the effect of a large value of σ^2 on the exact sample size is one of the most interesting issues in a sample size problem, we feel that this is a better formulation

than the alternative of putting, say, a noninformative prior on σ^2 . Using a prior on σ^2 will also lead to a different analytical problem and may be treated separately.

We derive the sample size for a variety of σ^2 s, keeping all other parameters in the problem fixed. Sample size is not expected to be robust with respect to drastic changes in σ^2 . If the exact value of σ^2 is unknown, a reasonable strategy is to choose n corresponding to the largest probable value of σ^2 . If no finite upper bound on σ^2 is available, the exact calculations that we are able to do would have to be replaced by purely numerical calculations after giving σ^2 a prior.

The matrix algebra of the problem involves only $k \times k$ matrices of the form $aI + bJ$, where I is the identity matrix, J is the $k \times k$ matrix of ones ($J = \underline{\mathbf{1}}\underline{\mathbf{1}}'$); and a and b are real numbers.

2.1 The problem

Let

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}; \quad 1 \leq i \leq k, \quad 1 \leq j \leq n, \quad (3)$$

be the observations for which

$$\epsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2). \quad (4)$$

We assume σ^2 is known. Additionally, let

$$\mu \sim N(0, \sigma_1^2)$$

and let

$$\tau = (\tau_1, \dots, \tau_k) \sim \pi_0 \mathbf{1}(\tau = 0) + \pi_1 MVN_k(0, \sigma_2^2 I)$$

be the priors on μ and τ , where $\pi_0(\pi_1)$ is the prior probability of hypothesis $H_0 : \tau = 0$ ($H_1 : \tau \neq 0$) ($\pi_0 + \pi_1 = 1$). We assume that μ and τ are independent. The choice of priors was made having the following two considerations in mind:

- (i) A point mass at $\tau = 0$ is needed since we want to test a precise null hypothesis, and
- (ii) The problem should stay mathematically manageable, so explicit sample size prescriptions can be made.

2.2 Bayes factor

Let $y = (\bar{y}_1, \bar{y}_2, \dots, \bar{y}_k)$, where $\bar{y}_i = \frac{1}{n} \sum_j y_{ij}$. The likelihood $f(\mu, \tau | y)$ for the above model is proportional to $\prod_{i=1}^k e^{-\frac{n}{2\sigma^2} (\bar{y}_i - \mu - \tau_i)^2}$, with a constant of proportionality depending on \bar{y}_i and σ^2 .

Let $\theta = (\mu, \tau)$.

Theorem 2.1 For testing

$$H_0 : \tau = 0 \quad v.s. \quad H_1 : \tau \neq 0,$$

the Bayes factor (in favor of H_1) is

$$B = \frac{|\Sigma_0|^{\frac{1}{2}}}{|\Sigma_1|^{\frac{1}{2}}} e^{-\frac{1}{2}y'(\Sigma_1^{-1} - \Sigma_0^{-1})y}, \quad (5)$$

where $\Sigma_0 = \frac{\sigma^2}{n}I + \sigma_1^2J$ and $\Sigma_1 = (\frac{\sigma^2}{n} + \sigma_2^2)I + \sigma_1^2J$.

Proof: The posterior probability of H_0 is

$$Pr(H_0|y) = \frac{\pi_0 \int_{\Theta_0} f(\theta|y) \mu_0(d\theta)}{\pi_0 \int_{\Theta_0} f(\theta|y) \mu_0(d\theta) + \pi_1 \int_{\Theta_1} f(\theta|y) \mu_1(d\theta)}, \quad (6)$$

where

$$\mu_0(\theta) = \mu_0(\mu, \tau) = N(0, \sigma_1^2) \times \mathbf{1}(\tau = 0); \quad \theta \in \Theta_0 = R \times \mathbf{0},$$

and

$$\mu_1(\theta) = N(0, \sigma_1^2) \times MVN_k(0, \sigma_2^2 I); \quad \theta \in \Theta_1 = R \times (R^k - \{0\}).$$

Let $\phi_\Sigma(x)$ denote the pdf of $MVN(0, \Sigma)$ distribution. Simple algebra gives that

$$Pr(H_0|y) = \frac{1}{1 + \frac{\pi_1}{\pi_0} B}, \quad (7)$$

where

$$B = \frac{\int_R \int_{R^k} \phi_{\frac{\sigma^2}{n}I}(y - \mu \mathbf{1} - \tau) \phi_{\sigma_2^2 I}(\tau) \phi_{\sigma_1^2}(\mu) d\mu d\tau}{\int_R \phi_{\frac{\sigma^2}{n}I}(y - \mu \mathbf{1}) \phi_{\sigma_1^2}(\mu) d\mu}. \quad (8)$$

The integral in the numerator of (8) is the marginal density of y when μ and τ are integrated out. Thus, the numerator is the density of $MVN_k(0, \Sigma_1)$ distribution, where $\Sigma_1 = (\frac{\sigma^2}{n} + \sigma_2^2)I + \sigma_1^2J$.

The integral in the denominator of (8) is the marginal of the conditional density of $y|\tau = 0$ when μ is integrated out. It is the density of $MVN_k(0, \Sigma_0)$ with $\Sigma_0 = \frac{\sigma^2}{n}I + \sigma_1^2J$.

The claim of the theorem now follows immediately. ■

2.3 Geometry of the problem

The following theorem describes the geometry of points in the observation space for which condition (1) cannot be satisfied.

Theorem 2.2 *The set K of y such that $\min\{Pr(H_0|y), Pr(H_1|y)\} > \epsilon$, can be described as:*

$$K = \{y \mid |y'(\Sigma_0^{-1} - \Sigma_1^{-1})y - \alpha| < \beta\} \quad (9)$$

where

$$\alpha = \log \frac{\pi_0^2 |\Sigma_1|}{\pi_1^2 |\Sigma_0|} = \log \frac{\pi_0^2 (\sigma^2/n + \sigma_2^2)^{k-1} (\sigma^2/n + k\sigma_1^2 + \sigma_2^2)}{\pi_1^2 (\sigma^2/n)^{k-1} (\sigma^2/n + k\sigma_1^2)}, \quad (10)$$

and

$$\beta = 2 \log \frac{1 - \epsilon}{\epsilon}. \quad (11)$$

Condition (2) on the probability content of the set K gives the relation from which the sample size can be found explicitly. *This explicit geometric description of the set K is extremely useful also due to the fact that when data are actually collected using the prescribed Bayesian sample size, whether or not the obtained data belong to K determines the validity of the prespecified posterior accuracy. Of course, one hopes that the obtained data will not belong to K .*

Proof: The condition

$$\epsilon < \min\{Pr(H_0|y), Pr(H_1|y)\} = \frac{1}{1 + \frac{\pi_1}{\pi_0} B} \mathbf{1}(B \geq \frac{\pi_0}{\pi_1}) + \frac{\frac{\pi_1}{\pi_0} B}{1 + \frac{\pi_1}{\pi_0} B} \mathbf{1}(B \leq \frac{\pi_0}{\pi_1}) \quad (12)$$

is equivalent to

$$\frac{\epsilon}{1 - \epsilon} \frac{\pi_0}{\pi_1} < B < \frac{1 - \epsilon}{\epsilon} \frac{\pi_0}{\pi_1}, \quad (13)$$

Since $\ln B = \ln \frac{|\Sigma_0|^{\frac{1}{2}}}{|\Sigma_1|^{\frac{1}{2}}} - \frac{1}{2} y'(\Sigma_1^{-1} - \Sigma_0^{-1})y$, by taking the logarithms in (13) we get (9). ■

Remark. Figure 1 depicts the posterior risk $\min\{Pr(H_0|y), Pr(H_1|y)\}$ as a function of Bayes factor B (relation (12)). We fixed $\pi_0 = 0.7$ and $\epsilon = 0.1$. Note that for $\frac{\pi_0}{\pi_1} = 7/3$ the risk is 0.5. The set of B s for which the risk exceeds $\epsilon = 0.1$ level corresponds to the set K in the observation space.

Geometrically, the set K is the region bounded by two nested k -dimensional ellipsoids of rotation, centered at the origin. The larger ellipsoid has one semiaxis equal to $\sqrt{\frac{\alpha + \beta}{\xi_1}}$, and $k - 1$ semiaxes equal to $\sqrt{\frac{\alpha + \beta}{\xi_2}}$, with $\xi_1 = \frac{n^2 \sigma^2}{(\sigma^2 + kn\sigma_1^2)(\sigma^2 + kn\sigma_1^2 + n\sigma_2^2)}$ and $\xi_2 = \frac{n^2 \sigma_2^2}{\sigma^2 + n\sigma_2^2}$ being the eigenvalues of $\Sigma_0^{-1} - \Sigma_1^{-1}$ of multiplicity 1 and $k - 1$, respectively. For the smaller ellipsoid the quantity $\alpha + \beta$ above is replaced by $\max\{\alpha - \beta, 0\}$.

It is important to demonstrate that the (marginal) probability content of the set K tends to zero when n goes to infinity.

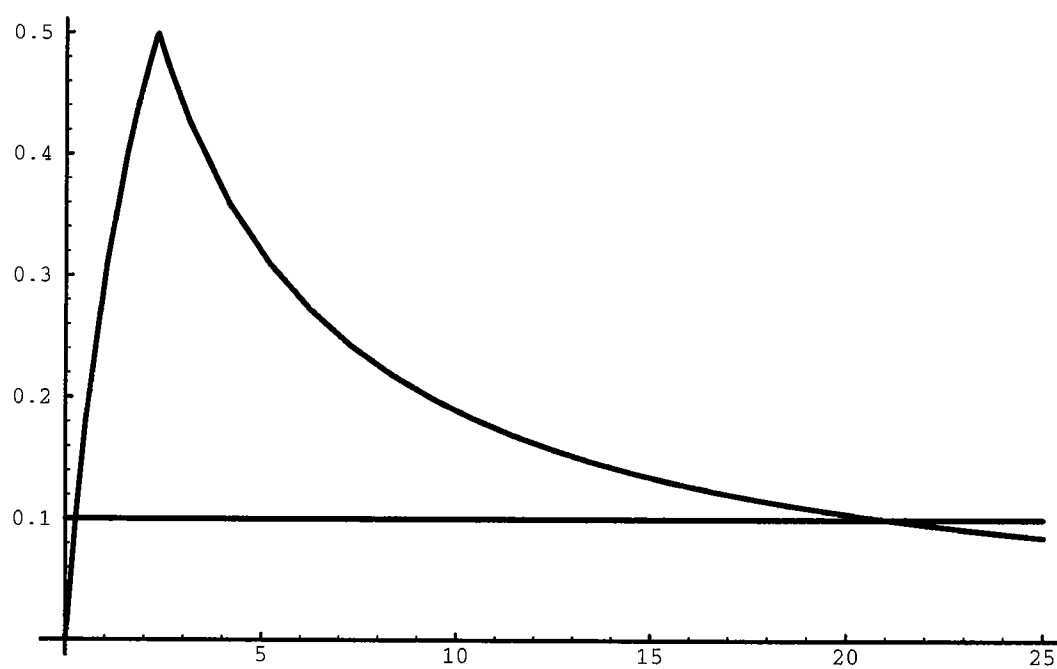


Figure 1: Posterior risk ($= \min\{P(H_0|y), P(H_1|y)\}$) against the Bayes factor, $\pi_0 = 0.7$, $\epsilon = 0.1$

Theorem 2.3 *Let P_n be a tight sequence of probability measures on \mathcal{B} , the Borel σ -algebra on \mathbb{R}^k , such that*

$$\frac{P_n(A)}{\ell(A)} \leq Q,$$

uniformly in n and in $A \in \mathcal{B}$, where $\ell(A)$ denotes the Lebesgue measure of the set A . Then

$$P_n(K) \rightarrow 0, \quad n \rightarrow \infty.$$

Proof: We give a proof for $k = 2$. For general k only minor modifications to the proof are needed. The set K , which also depends on n , is a subset of the ellipse E whose semiaxes have lengths equal to $\sqrt{\frac{\alpha+\beta}{\xi_2}}$ and $\sqrt{\frac{\alpha+\beta}{\xi_1}}$. The former behaves as $O(\sqrt{\log n})$ and the later as $O(\sqrt{\frac{\log n}{n}})$.

It is enough to show that $P_n(E) \rightarrow 0$, for $n \rightarrow \infty$. Let $\epsilon > 0$ be fixed. Since the family of probability measures P_n is tight one can find a ball B_ϵ of radius $\rho(B_\epsilon)$ such that $(\forall n)P_n(B_\epsilon) \geq 1 - \frac{\epsilon}{2}$. Then $P_n(E \cap B_\epsilon^c) \leq \frac{\epsilon}{2}$.

Since one axis of the ellipse E has length $O(\sqrt{\frac{\log n}{n}})$ one can find n_0 such that the length of the axis is less than

$$\frac{\epsilon}{2} \cdot \frac{1}{2\rho(B_\epsilon)} \cdot \frac{1}{2Q},$$

where Q is a constant that bounds the ratio $\frac{P_n(A)}{\ell(A)}$ from above uniformly in n and in $A \in \mathcal{B}$.

$$\text{Then } P_n(E \cap B_\epsilon) \leq Q \cdot \ell(E \cap B_\epsilon) \leq Q \cdot (2\rho(B_\epsilon)) \cdot 2\frac{\epsilon}{2} \cdot \frac{1}{2\rho(B_\epsilon)} \cdot \frac{1}{2Q} = \frac{\epsilon}{2}.$$

Thus for $n \geq n_0$,

$$P_n(K) \leq P(E) = P(E \cap B_\epsilon^c) + P(E \cap B_\epsilon) \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon. \quad \blacksquare$$

In our case the family P_n is a mixture of the tight families $MVN(0, \frac{\sigma^2}{n}I + \sigma_1^2J)$ and $MVN(0, (\frac{\sigma^2}{n} + \sigma_2^2)I + \sigma_1^2J)$. Any fixed mixture of two tight families is tight itself, which implies that the marginal family is tight. That the Lebesgue densities of P_n are uniformly bounded (in the variable and in n) also follows immediately from the mixture normal representation of P_n . This means that for *any* fixed ϵ and δ , a sample size satisfying the precision requirements can be found.

3 Equation for determining explicit sample size

The condition

$$Pr(|y'(\Sigma_0^{-1} - \Sigma_1^{-1})y - \alpha| \leq \beta) < \delta$$

is equivalent to

$$Pr(y'(\Sigma_0^{-1} - \Sigma_1^{-1})y \leq \alpha + \beta) - Pr(y'(\Sigma_0^{-1} - \Sigma_1^{-1})y \leq \alpha - \beta) < \delta. \quad (14)$$

To find n given the dimension k , σ^2 , σ_1^2 , σ_2^2 , ϵ and δ , such that (14) is satisfied, we need an effective way of calculating the probabilities of the type $Pr(y'Py < c)$.

The following is an adaptation of a result of Robbins and Pitman (1949). See also Johnson and Kotz (1970), pp 156-158.

Theorem 3.1 *Let P, Q be two $k \times k$ matrices of the form $aI + bJ$, $a, b \in R$ such that the following condition holds for the eigenvalues of the matrix $Q^{1/2}PQ^{1/2}$: The eigenvalue λ_1 of multiplicity 1 is smaller than the eigenvalue λ_2 of multiplicity $k - 1$.*

Let $y \sim MVN_k(0, Q)$. Then

$$Pr(y'Py < c) = \sum_{j=0}^{\infty} e_j Pr(\chi_{k+2j}^2 < c/\lambda_1), \quad (15)$$

where the coefficients e_i are obtained by the recurrence formula

$$e_0 = \left(\frac{\lambda_1}{\lambda_2}\right)^{\frac{k-1}{2}}$$

$$e_r = \frac{k-1}{2r} \sum_{j=0}^{r-1} e_j \left(1 - \frac{\lambda_1}{\lambda_2}\right)^{r-j}, \quad r \geq 1.$$

Proof:

The transformation $x = Q^{-1/2}y$, gives that (15) is equal to

$$Pr(x'Q^{1/2}PQ^{1/2}x < c) \stackrel{def}{=} \mathbf{F}(c|\lambda_1, \lambda_2) \quad (16)$$

with $x \sim MVN(0, I)$. λ_1 and λ_2 are eigenvalues of the matrix of the quadratic form.

The proof now follows from equation (41), page 157 in Johnson and Kotz (1977) by taking $\beta = \min(\lambda_1, \lambda_2)$. ■

Theorem 3.2

$$Pr(y \in K) = \pi_0(\mathbf{F}(c_2|\lambda_1^{(1)}, \lambda_2^{(1)}) - \mathbf{F}(c_1|\lambda_1^{(1)}, \lambda_2^{(1)})) + \quad (17)$$

$$\pi_1(\mathbf{F}(c_2|\lambda_1^{(2)}, \lambda_2^{(2)}) - \mathbf{F}(c_1|\lambda_1^{(2)}, \lambda_2^{(2)})) \quad (18)$$

where $c_1 = \alpha - \beta$, $c_2 = \alpha + \beta$,

$$\lambda_1^{(1)} = \frac{n\sigma_2^2}{\sigma^2 + kn\sigma_1^2 + n\sigma_2^2} \quad \text{and} \quad \lambda_2^{(1)} = \frac{n\sigma_2^2}{\sigma^2 + n\sigma_2^2}$$

are eigenvalues of $I - \Sigma_0^{1/2}\Sigma_1^{-1}\Sigma_0^{1/2}$, of multiplicity 1 and $k - 1$, respectively, and

$$\lambda_1^{(2)} = \frac{n\sigma_2^2}{\sigma^2 + kn\sigma_1^2} \quad \text{and} \quad \lambda_2^{(1)} = \frac{n\sigma_2^2}{\sigma^2}$$

are eigenvalues of $\Sigma_1^{1/2}\Sigma_0^{-1}\Sigma_1^{1/2} - I$, of multiplicity 1 and $k - 1$, respectively.

Proof: The marginal of y is $\pi_0MVN(0, \Sigma_0) + \pi_1MVN(0, \Sigma_1)$. Since $Pr(y \in K) = \pi_0Pr(y \in K|y \sim MVN(0, \Sigma_0)) + \pi_1Pr(y \in K|y \sim MVN(0, \Sigma_1))$, the result is a direct consequence of Theorem 3.1 applied on the two ellipsoids from equation (14) separately. ■

σ^2	n
0.5	31.52
0.7	44.13
1	63.04
1.5	94.56
2	126.05

Table 1: Some sample sizes

4 Examples and comparisons with classical sample sizes

The exact Bayesian sample size is determined by setting the expression (17) equal to δ . There are so many freely varying parameters that determine the exact value of n , that a comprehensive table of sample sizes is out of the question. Furthermore, computing technology has advanced to such a great extent that it is certainly more useful to provide a computer code for determining the sample size. A MATHEMATICA code is provided in Appendix. First we give an illustrative example by applying the computer algorithm to a specific case.

4.1 Exact sample sizes

Consider as an illustration the case when $\sigma_1^2 = \sigma_2^2 = 1, \pi_0 = 0.5, \epsilon = 0.1, \delta = 0.1$. The following table gives the sample sizes for different values of σ^2 when $k = 3$. Obviously, this is not intended to be a comprehensive table - it is only an artifact.

The table indicates that the Bayesian sample size may not be very robust with respect to specification of σ^2 . Since we already know that the sample size depends on the three variances only through the ratios σ_1^2/σ^2 and σ_2^2/σ^2 , the table equivalently indicates that careful elicitation of these two ratios is important, and not so much elicitation of the variances themselves.

Figure 2 gives level lines (in n) in an $\epsilon O \delta$ plane for the case of $\sigma_1^2 = \sigma_2^2 = \sigma^2, \pi_0 = 0.5$, and $k = 3$. This is an analog of very familiar similar sample size plots in the classical case. For example, $n = 20$ observations per treatment assure a posterior accuracy of 82% for all samples but for a set of predictive probability 0.18.

4.2 Comparison with classical sample sizes

As stated in the Introduction, it is interesting to investigate if using the widely available classical sample sizes themselves will meet the need of a Bayesian. Clearly, there are many reasonable ways to formulate and try to answer such a question. We will use the method outlined below which we think is very reasonable.

Step 1. Elicit $\sigma^2, \sigma_1^2, \sigma_2^2$ and π_0 .

Step 2. Fix a desirable level of posterior accuracy $1 - \epsilon$ and also a value of δ .

Step 3. Specify (reasonable) levels of type 1 and type 2 errors α, β . For determining the classical sample size it is also necessary to specify a measure of departure from the null hypothesis. Among many possible ways of doing so, it is quite common to specify a value

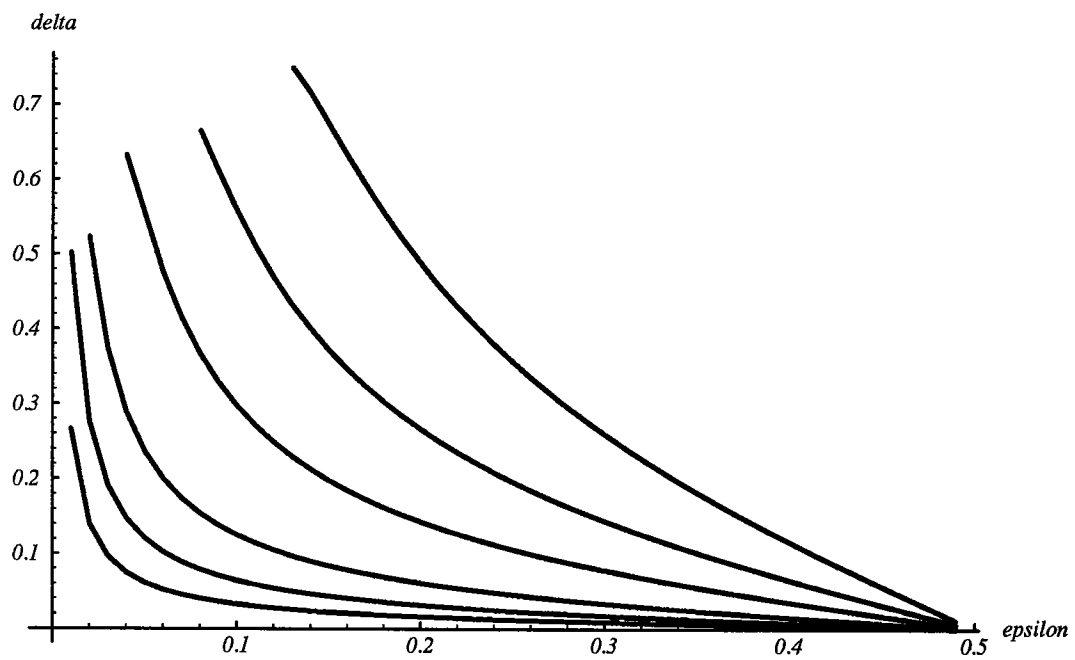


Figure 2: Level lines for $n = 5, 10, 20, 50, 100,$ and 200 listed from above.

k	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 1.5$
3	3	7	14
4	3	6	11
5	3	5	9
10	2	4	7

Table 2: Classical Sample Sizes (Step 3).

of

$$\Delta = \frac{\mu_{\max} - \mu_{\min}}{\sigma}, \quad (19)$$

where $\mu_i = \mu + \tau_i$.

Notice that Δ is a random variable in a Bayesian formulation. It will therefore be very natural to use the expected value of Δ under the prior as the specified value of Δ used to find the value of n from a table of classical sample sizes.

This expected value equals

$$\Delta_B = c_k \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sigma}, \quad (20)$$

where c_k equals the expectation of the range of a sample of size k from the standard normal distribution (for instance, $c_3 = 1.692$, $c_4 = 2.058$, $c_5 = 2.326$, and $c_{10} = 3.078$). This value can serve as a guideline even to classical statisticians for choosing Δ in sample size problems.

Step 4. For this particular value of n obtained in Step 3, identify the interval of values of σ for which the specified Bayesian precision requirement $Pr(K) \leq \delta$ is satisfied. Note that in the evaluation of $Pr(K)$ previously elicited values of σ_1 , σ_2 , and π_0 , and the prespecified value of ϵ are used. It is easy to see that this interval of values of σ is an interval of the form $\sigma \leq \sigma_0$.

Step 5. Compare σ_0 with the Bayesianly elicited value of σ . If the elicited value is unacceptably larger than σ_0 , it would be unreasonable to use the classical sample size and an independent derivation of the Bayesian sample size using the presented theory would be unavoidable. Otherwise, use of the classical sample size itself will suffice.

Here is an example.

Example.

Suppose, as an example, that the elicited values of σ_1^2 and σ_2^2 are both equal to 1 and π_0 is elicited to be 0.5. Suppose further that we want $\epsilon = 0.05$ and $\delta = 0.1$.

Table 2 gives the classical sample sizes for some selected values of k and using $\alpha = \beta = 0.05$. Recall that in this classical sample size calculation Δ is taken as in (20). The actual sample sizes are obtained from Harter and Owen (1975).

The elicited values of σ are 0.5, 1, and 1.5, respectively. Table 3 is self explanatory. Generally speaking the conclusion seems to be that the classical sample size should not be used if the number of treatments is small.

k	σ	σ_0	$\sigma \leq \sigma_0$	Conclusion
3	0.5	0.15756	no	do not use
	1	0.24067	no	do not use
	1.5	0.34037	no	do not use
4	0.5	0.27506	no	do not use
	1	0.38901	no	do not use
	1.5	0.52673	no	do not use
5	0.5	0.37347	no	do not use
	1	0.48524	no	do not use
	1.5	0.64685	no	do not use
10	0.5	0.56923	yes	use
	1	0.80501	no	do not use
	1.5	1.06493	no	do not use

Table 3

5 Appendix

In the Appendix we give a MATHEMATICA code for solving Bayesian sample size problem. For a given input $\sigma^2, \sigma_1^1, \sigma_2^2, k, n, \pi$, and ϵ the function ProContK calculates δ - the marginal probability content of the set K .

```

BeginPackage["Size`"]
Er[r_/;r==0, lam1_, lam2_, k_?Positive]:=
  Er[0, lam1, lam2, k]=
    (lam1/lam2)^((k-1)/2) //N;
Er[r_/;(r>0), lam1_, lam2_, k_?Positive]:=
  Er[r, lam1, lam2, k]=
    1/(2 r) Sum[ (k-1) (1- lam1/lam2)^(r-j) \
    Er[j, lam1, lam2, k], {j, 0, r-1}] //N;

QuadFormDist[arg_?Positive, lam1_, lam2_, \
  k_?Positive ]:=
Module[{s=0, ii=0, iter=-1},
ChiSquareDistribution/: CDF[ChiSquareDistribution[n_], x_] :=
  With[{result = N[GammaRegularized[n/2, 0, x/2]]},
If[NumberQ[result], result, GammaRegularized[n/2, 0, x/2]]];
  While[ s-iter > 10^(-10),
    iter = s;
    s += Er[ii, lam1, lam2, k]*
    CDF[ChiSquareDistribution[k+2 ii], arg/lam1]; \
    ii++
  ];
  s
];

ProContK[s_?Positive, s1_?Positive, \
  s2_?Positive, k_?Positive, \
  n_?Positive, pi0_?Positive, \
  eps_?Positive]:=
Module[{lam11, lam12, lam21, lam22, alpha, beta, \
  FF11, FF12, FF21, FF22}, \
  lam11= n s2/(s+k n s1 + n s2) //N;
  lam21= n s2/(s+ n s2) //N;
  lam12= n s2/(s+ k n s1)//N;
  lam22= n s2/s //N;
  alpha= Log[ pi0^2 (s/n+s2)^(k-1) (s/n+ k s1 + s2)/ \
    ( (1-pi0)^2 (s/n)^(k-1) (s/n + k s1) )]//N;
  beta= 2 Log[(1-eps)/eps];
  FF11=QuadFormDist[Max[ alpha-beta, 10^(-10)], lam11, lam21, k] //N;

```

```

FF21=QuadFormDist[ alpha+beta, lam11,lam21,k] //N;
FF12=QuadFormDist[Max[ alpha-beta, 10^(-10)], lam12,lam22,k] //N;
FF22=QuadFormDist[ alpha+beta, lam12,lam22,k] //N;
Return[ pi0 (FF21-FF11)+ (1-pi0) (FF22-FF12) ]
]
EndPackage[]

```

References

- [1] COCHRAN, W. (1962). *Sampling Techniques*. John Wiley & Sons, Inc.
- [2] CHALONER, K. (1984). Optimal Bayesian Experimental design for linear models. *Ann.Statist.* **12** 283-300.
- [3] DASGUPTA, A. AND MUKHOPADHYAY, S. (1994). Uniform and sub-uniform posterior robustness: The sample size problem. *Journal of Statistical Planning and Inference.* **40** *Special Issue on Bayesian Analysis, Ed. J. Berger, Number 2/3, 189-204.*
- [4] DASGUPTA, A. and STUDDEN, W. (1991). Robust Bayesian Experimental Designs in normal linear models. *Ann.Statist.* **19** 1244-1256.
- [5] HARTER, H. and OWEN, D. (1975). *Selected Tables in Mathematical Statistics*. Vol III, American Mathematical Society, Providence, RI.
- [6] HUTTON J. L. and OWENS, R. G. (1993). Bayesian sample size calculation and prior beliefs about child sexual abuse. *The Statistician* **42** 399-404.
- [7] JOHNSON, N. and KOTZ, S. (1970). *Continuous univariate distributions - 2*. J. Wiley & Sons, Inc
- [8] KASTENBAUM, M. A., HOEL, D. G., and BOWMAN, K. O. (1970). Sample size requirement: one-way analysis of variance. *Biometrika* **57** 421-430.
- [9] LEE, C., MUELLER, P., and PARMIGIANNI, G. (1993). Optimal sample size for two binomials. *Discussion Paper, 93-A25, ISDS, Duke University.*
- [10] ODEH, R. E. (1991). *Sample Size Choice : Charts for Experiments with Linear Models* M. Dekker.
- [11] PEARSON, E. S. and HARTLEY, H. O. (1951). Charts of the power function for Analysis of Variance tests derived from the non-central F-distribution. *Biometrika* **38** 112-130.
- [12] PILZ, J. (1991). *Bayesian estimation and Experimental Design in Linear Regression Models*. John Wiley & Sons, Inc.

- [13] ROBBINS, H. E. and PITMAN, E. J. G. (1949). Application of the method of mixtures to quadratic forms in normal variables. *Ann. Math. Statist.* **20** 552-560.
- [14] WOLFRAM, S. (1992). *Mathematica - A System for Doing Mathematics by Computer* - Second Edition, Addison-Wesley Publ. Comp.

DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
WEST LAFAYETTE, IN 47907
dasgupta@snap.stat.purdue.edu

INSTITUTE OF STATISTICS AND DECISION SCIENCES
DUKE UNIVERSITY
POBox 90251
DURHAM, NC 27708-0251
brani@isds.duke.edu