Bayesian Sample Size Calculations for Hypothesis Testing

Author(s): Robert Weiss

Source: *Journal of the Royal Statistical Society. Series D (The Statistician)*, 1997, Vol. 46, No. 2, Special Issue: Sample Size Determination (1997), pp. 185-191

Published by: Wiley for the Royal Statistical Society

Stable URL: https://www.jstor.org/stable/2988523

The Statistician (1997)
46, *No.* 2, *pp.* 185–191

Bayesian sample size calculations for hypothesis testing

By ROBERT WEISS†

University of California at Los Angeles, USA

[Received April 1996. Revised January 1997]

SUMMARY

A Bayesian approach to sample size calculation in hypothesis testing problems is developed. The sample size is chosen to make it *a priori* probable that the Bayes factor is greater than a given cut-off of prespecified size. Two methods for choosing a cut-off are given: an absolute criterion and a relative criterion. Calculations can be done using either exact algebraic manipulation or through simulation. The approach permits the propagation of uncertainty in quantities which are unknown and permits the computation of power and type I error rates either conditionally or unconditionally on particular values of the parameter of interest. A graphical tool is given for assessing the sensitivity of the predicted outcomes to model and sample size specification. The approach is illustrated for a one-sided and for a two-sided alternative hypothesis for continuous data with a normal prior.

Keywords: Bayes factor; Hypothesis test; Model selection

1. Introduction

The determination of sample size is of key importance in designing medical studies. Traditional classical methods of selecting sample sizes require a point specification of interesting treatment differences, variances and power and α-level. Classical methods are unable to take into account uncertainty in point specifications. Bayesian methods are ideally suited for design since they provide a language for specifying uncertainty and can properly propagate uncertainty, and elicitation is formally a part of the Bayesian paradigm. The Bayesian paradigm can be used to determine no treatment difference, something that cannot be done using current interpretations of the frequentist paradigm.

I consider the common situation of a medical study where a new treatment is being compared with a known control or standard. I assume that the researcher is interested in testing a null hypothesis H_0 against an alternative hypothesis H_1. Two common examples are

(a) a two-sided alternative where H_0 is no treatment effect *versus* the alternative hypothesis H_1 of a treatment effect and
(b) a one-sided alternative where H_0 is a negative or zero treatment effect *versus* H_1, a positive treatment effect.

Both the *marginal power* and the *conditional power* can be computed. Marginal power is defined as the marginal probability of concluding that H_1 is true, whereas the conditional power is the conditional probability of concluding H_1 given a particular treatment effect size. Both are potentially of interest; a clinical trial often has much value beyond proving that a particular treatment has an effect. Marginal power indicates the unconditional probability of rejecting a null hypothesis, whereas conditional power might be used to indicate that, if the treatment effect is sufficiently large to be of interest, the study has a good chance of identifying the existence of an effect. Both calculations might be of interest to a funding agency.

†*Address for correspondence*: Department of Biostatistics, School of Public Health, University of California at Los Angeles, Los Angeles, CA 90095-1772, USA.
E-mail: rob@rem.ph.ucla.edu

 0039–0526/97/46185

This paper develops the log-Bayes-factor $\log B$ as the test statistic when prior probabilities of the two models are not used. Previous work has used Bayesian methods to compute the power for a classical hypothesis test (Brown *et al.*, 1987; Spiegelhalter and Freedman, 1986; Freedman and Spiegelhalter, 1983), or to minimize the length of a posterior probability interval or to guarantee minimum posterior coverage for a given length (Joseph *et al.*, 1995; Adcock, 1993; Hutton and Owens, 1993; Pham-Gia and Turkkan, 1992). Another approach to design uses sequential trials and a stopping rule that minimizes the cost of continuing in terms of dollars or patients' lives (Berry and Ho, 1988; Sylvester, 1988). Korn (1990) described some pitfalls in using maximum likelihood point estimates from a previous study to estimate power for a future study; in contrast, the current work uses a prior distribution from a previous study or other source to estimate the sample size for the future study.

The paper is structured as follows. The next section gives the basic methodology. Sections 3 and 4 illustrate the key ideas in two-sided and one-sided hypothesis testing situations for a normal mean. The paper closes with a brief discussion.

2. Methodology

The standard Bayesian mechanism for testing two hypotheses is to compute the posterior probability of either hypothesis given the data, the model, the study design and the priors under each model. Often it is not desired to specify a prior probability for each hypothesis, and the inference is based on the Bayes factor $B(0\colon 1|Y)$ in favour of H_0 against H_1:

$$B(0\colon 1|Y) = \frac{f(Y|H_0)}{f(Y|H_1)},$$

where

$$f(Y|H_j) = \int f(Y|\theta_j)\, p(\theta_j|H_j)\, \mathrm{d}\theta_j,$$

and where Y are the data, θ_j are the parameters under hypothesis H_j, $j = 0, 1$, $f(Y|\theta_j)$ is the sampling distribution of Y under H_j and $p(\theta_j|H_j)$ is the prior under the hypothesis H_j. Each hypothesis may have a separate sampling distribution and prior, although the typical situation will be that the sampling distributions are the same and the two hypotheses differ only in the specification of the prior. Since $B(0\colon 1|Y)$ is non-negative, it is convenient to take logarithms:

$$b_{01} \equiv b(0\colon 1|Y) = \log B(0\colon 1|Y).$$

A positive value of b_{01} means that the data support H_0, and a negative value means that the data support H_1. Kass and Raftery (1995) give a summary of the current state of Bayes factors. They suggest interpreting b_{01} between 3 and 5 as strong evidence in favour of H_0 and $b_{01} > 5$ as very strong evidence. Negative b_{01} of the same magnitude favour H_1 by the same amount.

Bayesian sample size selection might aim for a reasonable probability that b_{01} be greater than or less than a prespecified non-negative number k or $-k$ under H_0 and H_1 respectively. Alternatively the probabilities $P(b_{01} > k|H_0)$ and $P(b_{01} < -k|H_1)$ averaged across H_0 and H_1 might be, say, 0.5. The type I error would be defined as $P(b_{01} \leqslant k|H_0)$ and the type II error would be $P(b_{01} > -k|H_1)$. One approach, for example, might take 0 as the cut-off and require that $P(b_{01} > 0|H_0)$ and $P(b_{01} < 0|H_1)$ be quite high. The sample size would be chosen to meet constraints on these two errors. Alternatively, a cut-off of $k = 3$ or $k = 5$ could be chosen, and we might require that $P(b_{01} > k|H_0)$ and $P(b_{01} < -k|H_1)$ be approximately 0.5.

Another approach to specifying a cut-off borrows from the classical paradigm. The cut-off could be part of the design. For example, the cut-off could be chosen on the basis of the distribution of b_{01} under hypothesis H_0; in particular, we might take $b_{01,\alpha}$, $0 < \alpha < 1$, the $100\alpha\%$ lower tail of $p(b_{01}|H_0)$ as the cut-off k; n would be chosen to make $p(b_{01} < b_{01,\alpha}|H_1)$

occur with some high prespecified probability $1-\beta$. We might choose $\alpha = 0.05$ or $\alpha = 0.01$ and $\beta = 0.2$ or $\beta = 0.1$. I call this a *relative* choice of cut-off. Partial motivation for this treatment is that it is less stringent and therefore requires a smaller sample size than the absolute cut-off with, say, $k = 5$. If the cut-off $b_{01,\alpha}$ must be calculated using simulation as will often be the case, this procedure illustrates designs for Monte Carlo tests of hypotheses (Diggle and Gratton, 1984; Besag and Clifford, 1989).

Because the data have not yet been observed, the data are random. The Bayes factor, a function of the data, is also random. Since the Bayes factor is a univariate quantity, the two distributions $p(b_{01}|H_0)$ and $p(b_{01}|H_1)$ can be plotted on a graph and inspected as any and all design parameters are modified. These two distributions show the expected outcomes of the study conditionally on each hypothesis. The plot shows how well separated the two hypotheses are. Distributions with substantial overlap indicate hypotheses that cannot be separated, possibly because of an inappropriate design or because of limited sample size. Hopefully, but not necessarily guaranteed, as the example in the next section shows, we would find $P(b_{01} > k|H_0)$ and $E[b_{01}|H_0]$ increasing at a reasonable rate with increasing sample size and similarly $P(b_{01} < -k|H_1)$ and $E[b_{01}|H_1]$ decreasing with increasing sample size.

3. Sample size selection for precise null hypothesis

Possibly the simplest situation is testing whether a normal mean is equal to 0 or not. Assume that $y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$. Under hypotheses H_j, $j = 0, 1$, assume that $\mu \sim N(\mu_j, \tau_j)$ with σ^2, μ_j and τ_j all known. The values for μ_j and τ_j could be determined by elicitation, literature review or from pilot data. Without loss of generality, take $\mu_0 = 0$. Taking H_0 to be a precise null hypothesis gives $\tau_0 = 0$. This gives a common choice of prior for H_0. This example is already of practical interest, since y_i can be the difference in outcomes of paired observations, or we could be comparing the treatment with an established standard.

Our goal is to select a sample size n so that the two hypotheses can be distinguished. The classical sample size n_c would presumably set

$$n_c = \frac{\sigma^2(z_\alpha + z_\beta)^2}{\mu_1^2},$$

where α is the type I error, often 0.05 or 0.01, or 0.025 or 0.005 for two-sided tests, and the power $1-\beta$, often 0.8 or 0.9, is 1 minus the type II error and $\Phi(z_\alpha) = 1-\alpha$ and $\Phi(z_\beta) = 1-\beta$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function. This formula for n_c assumes that the point estimate μ_1 of μ under H_1 is the difference of interest. The classical sample size does not involve τ_1, which seems odd. Frequently, the choice of $\mu_0 = 0$ is a nominal value to illustrate the hypothesis of no difference, whereas the prior distribution under H_1 is based on quantifiable prior information. If $\tau_1^{1/2}$ is very small compared with μ_1, μ is already well determined and a new study is not needed.

Another, Bayesian, methodology is to have the posterior $1-\alpha$ probability content interval have width w. This methodology might select n to be the smallest integer greater than $(4w^{-2}z_{\alpha/2}^2 - \tau_1^{-1})\sigma^2$. If this were negative, then no sample would be taken.

The predictive density of $Y = (y_i)$ under H_1 is

$$f(Y|H_1) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{(n-1)/2}(n\tau_1 + \sigma^2)^{0.5}} \exp\left[-0.5\left\{\frac{\mathrm{RSS}}{\sigma^2} + \frac{(\bar{y}-\mu_1)^2}{\sigma^2/n + \tau_1}\right\}\right]$$

where $\mathrm{RSS} = \Sigma(y_i - \bar{y})^2$ and $n\bar{y} = \Sigma y_i$. This formula also applies to $f(Y|H_0)$ by setting $\mu_1 = 0$ and $\tau_1 = 0$. Then

$$b_{01} = 0.5\left\{\log\left(\frac{n\tau_1 + \sigma^2}{\sigma^2}\right) + \frac{(\bar{y}-\mu_1)^2}{\sigma^2/n + \tau_1} - \frac{n\bar{y}^2}{\sigma^2}\right\}$$

$$= 0.5\left\{\log\left(1+\frac{n\tau_1}{\sigma^2}\right)+\frac{\mu_1^2}{\tau_1}-\frac{\tau_1}{(\sigma^2/n)(\sigma^2/n+\tau_1)}\left(\bar{y}+\frac{\mu_1\sigma^2}{n\tau_1}\right)^2\right\}. \quad (1)$$

The distributions of interest are $p(b_{01}|H_j)$. The RSS term conveniently drops out of equation (1), which means that in evaluating the distribution of b_{01} under both models we need only be concerned with the prior predictive distribution of $\bar{y}|H_j$.

First I illustrate the relative cut-off approach for specifying the sample size. For a given sample size n, I find the cut-off value $k = b_{01,\alpha}$, and the probability $P(b_{01} > k|H_1)$, then I search for n to make $P(b_{01} > k|H_1) = 1-\beta$. Inspection of equation (1) shows that b_{01} can be written as

$$b_{01} = A - B(\bar{y}+C)^2,$$

where A, B and C are all positive, and

$$A = 0.5\left\{\log\left(1+\frac{n\tau_1}{\sigma^2}\right)+\frac{\mu_1^2}{\tau_1}\right\},$$

$$B = 0.5\frac{\tau_1}{(\sigma^2/n)(\sigma^2/n+\tau_1)},$$

$$C = \frac{\mu_1\sigma^2}{n\tau_1}.$$

Under hypothesis H_j, $j = 0, 1$, $\bar{y}|H_j \sim N(m_j, s_j^2)$, with $m_0 = 0$, $s_0^2 = \sigma^2/n$, $m_1 = \mu_1$ and $s_1^2 = \sigma^2/n + \tau_1$. The probability $P(b_{01} < A|H_j) = 1$ for either hypothesis, so, if $k > A$, $P(b_{01} < k|H_j) = 1$, and A is an upper bound for b_{01} under both H_0 and H_1. Assuming $k < A$ and for a fixed sample size n, straightforward but tedious algebraic manipulation shows

$$P(b_{01} < k|H_j) = \Phi\left\{-\left(\frac{A-k}{Bs_j^2}\right)^{1/2}-\frac{C+m_j}{s_j}\right\}+1-\Phi\left\{\left(\frac{A-k}{Bs_j^2}\right)^{1/2}-\frac{C+m_j}{s_j}\right\}. \quad (2)$$

Using equation (2), and setting $P(b_{01} < b_{01,\alpha}|H_0) = \alpha$ for given α, one can use root-finding software to find the α-quantile for b_{01} under H_0 for a given sample size. Then, given $k = b_{01,\alpha}$, one uses the same formula to calculate $P(b_{01} < b_{01,\alpha}|H_1)$. With a particular choice of α such as $\alpha = 0.05$ or $\alpha = 0.01$, various techniques can be used to search for the sample size n which gives $P(b_{01} < b_{01,\alpha}|H_1) = 1-\beta$ for a prespecified $\beta = 0.1$ or $\beta = 0.2$.

To illustrate a particular example, consider the situation where $\tau_1 = 1$, $\sigma^2 = 5^2$ and $\mu_1 = 2$, and the classical difference of interest is 2. I take $\alpha = 0.05$ and $\beta = 0.2$, i.e. the type I error is 0.05 and the power is 0.8. The usual classical sample size formula for this situation gives $n = 5^2(1.96+0.84)^2/2^2 = 49.0555$ which would be rounded up to a sample of size 50. The sample size to obtain a posterior content of 0.95 in an interval of width 2 is $n = 24$.

Here I use the relative approach to specify the sample size. Increasing sample sizes $n = 20$, 40, 60 were checked to bracket the desired sample size. At $n = 60$, the value 65 was tried followed by 66 and 67; at $n = 66$, the power is slightly under 0.8. Table 1 summarizes the results of the calculations performed. Table 1 also gives the type I error and power using 0 as a cut-off and gives the conditional type I error and conditional power against the alternative point hypothesis H_1: $\mu = 2$ using either $b_{01,0.05}$ or 0 as the cut-off.

Fig. 1 plots $p(b_{01}|H_j)$, with $j = 0$ and $j = 1$ and for three sample sizes $n = 20$, 40, 60. The three densities on the right-hand side are for hypothesis H_0, with n increasing from left to right. The three densities on the left-hand side are for H_1 with n increasing from right to left. As n increases, the two densities $p(b_{01}|H_j)$ separate. We see two important features in this plot. First, under H_0, it is not possible to obtain a large Bayes factor in favour of H_0. This

TABLE 1
$\alpha = 0.05$ quantile of the distribution $p(b_{01}|H_0)$, the probability $p(b_{01} < b_{01,0.05}|H_1)$ or power, the type I error α and power using 0 as the cut-off, the conditional type I error α and power against the point hypothesis H_1: $\mu = 2$, using the cut-off and the conditional type I error rate and power using 0 as a cut-off, for various sample sizes

Sample size	*Unconditional* H_1: $\mu \sim N(2, 1)$		*Conditional versus* H_1: $\mu = 2$					
	Cut-off	*Power*	$b_{cut} = 0$		$b_{cut} = b_{01,\alpha}$		$b_{cut} = 0$	
			α	*Power*	α	*Power*	α	*Power*
20	−1.05	0.543	0.16	0.73	0.069	0.62	0.19	0.81
40	−0.72	0.708	0.10	0.79	0.060	0.84	0.10	0.90
60	−0.43	0.785	0.076	0.82	0.046	0.92	0.06	0.94
65	−0.37	0.797	0.072	0.82	0.042	0.93	0.05	0.95
66	−0.35	0.800	0.071	0.82	0.041	0.94	0.05	0.95
67	−0.34	0.802	0.070	0.83	0.041	0.94	0.05	0.95

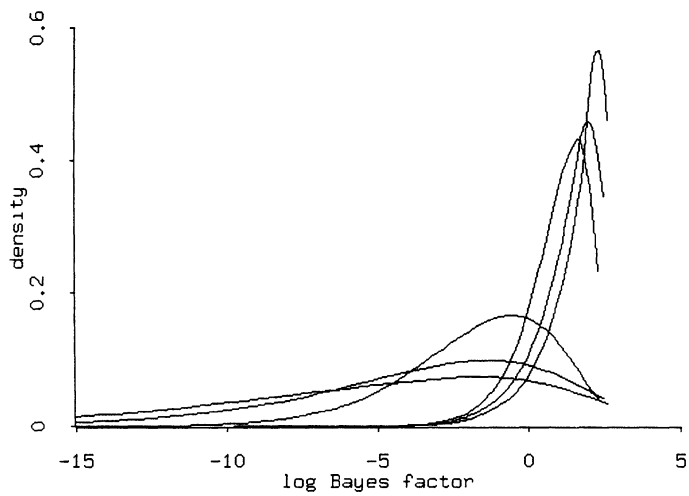

Fig. 1. Plots of $p(b_{01}|H_j)$ for $j = 0$, 1 and for sample sizes $n = 20$, 40, 60: the densities are kernel density estimates based on samples of size 25000 from the respective densities; as the sample size increases the density of $b_{01}|H_0$ moves only slightly to the right with increasing sample size, but increasing sample size does affect the density under hypothesis H_1

persists over all reasonable choices of n and is in fact how I first realized that b_{01} has a maximum value which varies extremely slowly with n. Originally I wished to have a sample size which gave a substantial probability of obtaining $b_{01} > 5$ or $b_{01} < -5$ under hypotheses H_0 and H_1 respectively. However, this would require an enormous sample size for the given priors. For example, a sample of size 67 has a maximum value for b_{01} of only 2.65. A sample of size 10061 is needed to give a maximum value of 5. Requiring a high probability of a large Bayes factor is helpful so that minor perturbations to the data do not affect the general conclusions and so that larger variations in the prior beliefs of the audience are overridden by the data. The relative approach demonstrated here worked well and gave reasonable sample sizes. An absolute approach based on even moderate probabilities that the Bayes factor be

larger than 5, or other similar value of the cut-off k, will not work. One possible compromise is to relax the requirement of a large positive log-Bayes-factor under H_0 but still to require a reasonable probability of a large negative b_{01} under H_1. A second reasonable compromise is to take the cut-off to be 0. Table 1 gives power for this cut-off for the sample sizes indicated. For example, a sample size of 65 gives probabilities of 0.93 and 0.82 under H_0 and H_1 of having the Bayes factor of the correct sign under each hypothesis.

The second feature that we see in Fig. 1 is that under hypothesis H_1 there is substantial probability that b_{01} will be very large and negative. For a sample of size 67, there is a probability $P(b_{01} \leqslant -10|H_1) = 0.23$. We may not wish to continue sampling when the evidence is this strong against the null hypothesis. This suggests a consideration of multistage procedures.

Simulation could have easily been used to do the calculations in this section. Basically, simulation would have prevented the need for any algebra after equation (1), with a cost that the calculations would have been less accurate. In more complicated design problems, exact calculation is not possible, and simulation is mandatory. Fig. 1, for example, plots kernel density estimates based on samples of size 25000 from the relevant densities. Doing these calculations in closed form would require differentiating equation (2) with respect to k. Simulation was used to check the results in Table 1 for the second to fifth columns. The simulation results agreed with the exact results up to simulation error.

4. One-sided hypothesis

Here I briefly consider a one-sided alternative hypothesis. I consider the same prior information as before, $\mu \sim N(\mu_1, \tau_1)$, but take H_0: $\mu \leqslant 0$ and H_1: $\mu > 0$. The prior probability of H_0 is then $\Phi(-\mu_1/\tau_1^{1/2})$, and this value is 0.023 for our numerical example and the prior probability of H_1 is 0.977. Monte Carlo calculations based on samples of size 25000 were used for the remaining calculations reported here; both algebraic and computational details have been omitted. Choosing $\alpha = 0.05$ and $\beta = 0.2$, and choosing n to achieve a relative separation of the distributions of b_{01} under the two hypotheses, required a sample size of 40, with an estimated power of 0.801. A sample size of 39 gave an estimated power of 0.78. A sample of size $n = 50$ was needed to satisfy a design specification requiring 0.5 probability of having $b_{01} > 5$ under hypothesis H_0 or $b_{01} < -5$ under H_1 averaged by the prior probabilities of H_0 and H_1.

5. Discussion

Both conditional and unconditional power are of interest. Unconditional power is the marginal prior probability of discovering that the new treatment is different from the old treatment. Conditional power calculates the probability of discovering that a difference exists given that the difference is a certain size. Most clinical trials study much more than just the treatment effects and much information is gained about the course of disease, the satisfaction of patients, details of treatment and care. Funding should be partially justified on the basis of this ancillary information with the further requirement that a clinically useful (or likely) difference in outcomes due to treatment be findable with high power.

More standard statistics such as Wald statistics may be studied by using the same methodology. In principle, dosage or other settable treatment parameter specifications can be designed using this methodology although additional computational methods would need development. Robustness studies might generate data under a model that is different from the likelihood used in the analysis. It is also possible to consider using different densities for the prior of θ for generating θ and in calculating the predictive $f(Y|M_j)$. This would happen if a nominal vague prior were to be used in place of the informative prior used to generate θ and the data.

Acknowledgement

This work was supported by grant GM 50011 from the National Institute of General Medicine of the United States National Institutes of Health.

References

Adcock, C. J. (1993) An improved Bayesian procedure for calculating sample sizes in multinomial sampling. *Statistician*, **42**, 91–95.

Berry, D. A. and Ho, C. H. (1988) One-sided sequential stopping boundaries for clinical trials: a decision-theoretic approach. *Biometrics*, **44**, 219–227.

Besag, J. and Clifford, P. (1989) Generalized Monte Carlo significance tests. *Biometrika*, **76**, 633–642.

Brown, B. W., Herson, J., Atkinson, E. N. and Rozell, M. E. (1987) Projection from previous studies: a Bayesian and frequentist compromise. *Contr. Clin. Trials*, **8**, 29–44.

Diggle, P. J. and Gratton, R. J. (1984) Monte Carlo methods of inference for implicit statistical models (with discussion). *J. R. Statist. Soc.* B, **46**, 193–227.

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

Hutton, J. L. and Owens, R. G. (1993) Bayesian sample size calculations and prior beliefs about child sexual abuse. *Statistician*, **42**, 399–404.

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

Kass, R. E. and Raftery, A. (1995) Bayes factors. *J. Am. Statist. Ass.*, **90**, 773–795.

Korn, E. L. (1990) Projecting power from a previous study: maximum likelihood estimation. *Am. Statistn*, **44**, 290–292.

Pham-Gia, T. and Turkkan, N. (1992) Sample size determination in Bayesian analysis. *Statistician*, **41**, 389–397.

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

Sylvester, R. J. (1988) A Bayesian approach to the design of Phase II clinical trials (C/R: V46 p535–538). *Biometrics*, **44**, 823–836.