Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713428038

Geometric sample size determination in Bayesian analysis

M. M. Nassar [a]; S. M. Khamis [a];S. S. Radwan [b]
[a] Department of Mathematics, Faculty of Science, Ain Shams University, Abbassia, Cairo, Egypt [b] Department of Mathematics, Faculty of Science, Al Azhar University, Girls Section, Nasr City, Cairo, Egypt

First published on: 03 March 2010

To cite this Article Nassar, M. M. , Khamis, S. M. andRadwan, S. S.(2010) 'Geometric sample size determination in Bayesian analysis', Journal of Applied Statistics, 37: 4, 567 — 575, First published on: 03 March 2010 (iFirst)
To link to this Article: DOI: 10.1080/02664760902803248
URL: http://dx.doi.org/10.1080/02664760902803248

Journal of Applied Statistics
Vol. 37, No. 4, April 2010, 567–575

Taylor & Francis
Taylor & Francis Group

Geometric sample size determination in Bayesian analysis

M.M. Nassar[a], S.M. Khamis[a]* and S.S. Radwan[b]

[a]*Department of Mathematics, Faculty of Science, Ain Shams University, Abbassia, Cairo, Egypt;*
[b]*Department of Mathematics, Faculty of Science, Al Azhar University, Girls Section, Nasr City, Cairo, Egypt*

(*Received 24 February 2008; final version received 27 January 2009*)

The problem of sample size determination in the context of Bayesian analysis is considered. For the familiar and practically important parameter of a geometric distribution with a beta prior, three different Bayesian approaches based on the highest posterior density intervals are discussed. A computer program handles all computational complexities and is available upon request.

Keywords: Bayesian analysis; average coverage criterion (ACC); average length criterion (ALC); worst-outcome criterion (WOC)

1. Introduction

The starting point for experimentation in the classical approach to statistics is frequently choosing the appropriate sample size. In the Bayesian context, however, it is important to find the suitable sample size prior to sampling, mainly due to time and cost constraints. Sample size determination arises in many statistical problems such as Bayesian analysis, decision theory and quality management.

Let p be an unknown probability of which inference is about to be done by taking a sample of size n. This problem has been treated by several authors. Adcock [1,3–5], Joseph et al. [7], Pham-Gia [9] and Pham-Gia and Turkkan [10] have treated binomial or multinomial sampling where the response variable is dichotomous (takes values {0, 1} only). Adcock [2] has treated the case of a normally distributed response variable.

Gelfand and Wang [6] have been concerned with the selection of sample size by adopting a screening criterion. Pezeshk [8] has reviewed some key techniques of Bayesian methods of sample size determination. In clinical trials, the determination of sample size under normal likelihoods has been discussed by Sahu and Smith [12] where at substantive testing stage of financial audit, they concluded that normality is not an appropriate assumption. Bayesian approaches have been formulated by Stamey et al. [14] to the problem of determining sample size for Bayesian interval

*Corresponding author. Email: soheir_khamis@hotmail.com

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664760902803248
http://www.informaworld.com

estimators of a predetermined length for a single Poisson rate, for the difference between two Poisson rates and for the ratio of two Poisson rates. A Bayesian simulation approach has been developed by Stamey and Gerlach [13] for determining the sample size required for a case–control study with misclassified data using an average posterior variance criterion considering both fixed cost and fixed variance procedures.

The purpose of this paper is to treat the problem of determining the sample size in the case of a geometric response variable. Some of the most common examples of such behavior are clustering of plants, bacterial clustering, etc., which provide an excellent model when other models are considered inapplicable.

Consider now a very important example in measuring patterns of plant populations. To investigate the relationship between species of plants growing together in an area, it has been suggested that their degree of 'segregation' be determined. Two species are said to be 'unsegregated' if the individuals of each are randomly intermingled with those of the other; if not, they are 'segregated'. In detecting and measuring segregation, one may wish to count the number of plant individuals of one species growing in the spaces among the clumps of the other species. The 'spaces' sampled are thus of variable extent; they are bounded, not by the edges of a ground frame, but by individuals of the second species [11]. Denoting the probability of encountering individuals of some species by $(1-p)$, the probability of obtaining an uninterrupted sequence (or run) of x_i, $x_i = 1, 2, \ldots$ individuals of first species is $(1-p)^{x_i-1}p$; likewise, the probability of obtaining a run of y_j, $y_j = 1, 2, \ldots$ of second species is $p^{y_j-1}(1-p)$ or $(1-q)^{y_j-1}q$, where $q = 1-p$. So for both species, the distribution of run lengths is geometric.

Now, suppose that the parameter p of the geometric distribution is itself a random variable having a constant value within any one run, but different values in different runs. Since $0 < p < 1$, one choice, in this case, for the distribution of p is to assume that it is a $\beta\,(c, d)$, i.e.

$$g(p) = \frac{1}{\beta(c,d)}\, p^{c-1}(1-p)^{d-1}.$$

Thus obtaining

$$\begin{aligned} P(\text{a run of } n \text{ individuals}) &= \frac{1}{\beta(c,d)} \int_0^1 p^{c+n-2}(1-p)^d \mathrm{d}p \\ &= \frac{\beta(c+n-1, d+1)}{\beta(c,d)}. \end{aligned}$$

Another example is the count of the number of newborns having the specific congenital malformation in groups of consecutive births. The number of consecutive births occurring between the birth of an infant with the specific malformation is being monitored and the birth of the next infant with that malformation. Such a group of consecutive births is defined as a run and its length is a geometric variate. The parameter of the geometric random variable, as above, is itself a random variable having a beta distribution.

2. Bayesian criteria for sample size

Let 'n' denote the sample size, 'p' the parameter of interest, Ω the parameter space for p and $g(p)$ the prior distribution of p. The experiment will consist of observing n data points $x = (x_1, x_2, \ldots, x_n)$, where x is composed of n exchangeable components from the data space X. The pre-posterior marginal distribution of x is

$$f(x) = \int f(x|p)g(p)\mathrm{d}p, \tag{1}$$

and the posterior distribution of p given data x is

$$\Psi(p|x,n) = \frac{f(x|p)g(p)}{\int_\Omega f(x|p)g(p)\mathrm{d}p}, \tag{2}$$

where $f(x|p)$ is the likelihood of the data.

As mentioned by Joseph et al. [7], if p is one-dimensional and $\Psi(p|x,n)$ is unimodal, (a,b) is a highest posterior density (HPD) interval if and only if $\Psi(p_1|x,n) \geq \Psi(p_2|x,n)$ for all p_1 in (a,b) and all p_2 not in (a,b), i.e. the corresponding condition for (a,b) to be the HPD region is

$$\int_a^b \psi(p|x,n)\,\mathrm{d}p = 1-\alpha.$$

The methods considered suggest sample sizes that satisfy criteria relating in some way to either the variance of the posterior distribution for p or posterior coverage probabilities for intervals of prespecified length.

2.1 *Average coverage criterion*

For a given fixed HPD interval length ℓ, find the minimum sample size n such that the expected coverage probability of x is at least $1-\alpha$, *i.e.* the smallest n satisfying

$$\sum_x \left\{ \int_{a(x,n)}^{a(x,n)+\ell} \Psi(p \mid x,n)\mathrm{d}p \right\} f(x) \geq 1-\alpha, \tag{3}$$

where $a(x,n)$ is the lower limit of the HPD credible set of length ℓ for the posterior density $\Psi(p \mid x,n)$.

2.2 *Average length criterion*

An alternative way to select a sample size would be to fix the coverage probability $1-\alpha$ of the HPD credible set for p. Then each possible outcome x will require a certain length $\ell'(x,n)$ to obtain the desired coverage probability. This criterion ensures that, for any x in space X,

$$\int_{a(x,n)}^{a(x,n)+\ell'(x,n)} \Psi(p \mid x,n)\mathrm{d}p = 1-\alpha, \tag{4}$$

and the problem is to select the sample size according to the smallest n such that the expectation (with respect to x) of these lengths is less than ℓ, i.e. the smallest n that satisfies

$$\sum_x \ell'(x,n) f(x) \leq \ell, \tag{5}$$

where ℓ is the pre-specified average length that is desired.

2.3 *Worst-outcome criterion*

Both of the above criteria give no guarantee for any particular observed data x. The most conservative approach would be to ensure that the requirements for both length and coverage probability hold simultaneously over all possible data x. Thus, rather than averaging, a minimum n is chosen such that

$$\inf_{x\in X} \left\{ \int_{a(x,n)}^{a(x,n)+l} \Psi(p \mid x,n)\mathrm{d}p \right\} \geq 1-\alpha, \tag{6}$$

where both ℓ and α are fixed in advance.

3. Bayesian sample size for geometric proportions

This section applies the pre-discussed criteria to the estimation of p, the probability of a run in geometric sampling.

Consider this probability p as a random variable with a beta prior distribution:

$$g(p; c, d) = \frac{1}{\beta(c, d)} p^{c-1}(1-p)^{d-1}, \tag{7}$$

where $0 < p < 1$, $c, d > 0$.

Considering geometric sampling for x, we have the probability function of the data:

$$f(x_i \mid p) = p(1-p)^{x_i}, \quad x_i = 0, 1, 2, \ldots. \tag{8}$$

With the data points $x^* = (x_1, \ldots, x_n)$ and $x = \sum_{i=1}^{n} x_i$, the distribution of x is given by

$$f(x) = \frac{\beta(n+c, x+d)}{\beta(c, d)}. \tag{9}$$

Then the posterior distribution of p is also of beta type and is given by:

$$\Psi(p \mid n, c, d, x) = \frac{1}{\beta(n+c, x+d)} p^{n+c-1}(1-p)^{x+d-1}, \quad 0 < p < 1. \tag{10}$$

The means and variances of the prior and posterior distributions of p are then, respectively,

$$\mu_{\text{prior}} = \frac{c}{c+d}, \quad \text{var}_{\text{prior}} = \frac{cd}{(c+d)^2(c+d+1)}, \tag{11}$$

and

$$\mu_{\text{post}} = \frac{n+c}{n+c+d+x}, \quad \text{var}_{\text{post}} = \frac{(n+c)(x+d)}{(n+c+d+x)^2(n+c+d+1)}. \tag{12}$$

Now, applying the three criteria discussed in Section (2) to the estimation of the probability p, we have the following.

3.1 *Average coverage criterion*

In the context of a geometric parameter, from Equations (3), (9) and (10), we can say average coverage criterion (ACC) then reduces to finding the minimum n that satisfies

$$\sum_x \frac{1}{\beta(c, d)} \left\{ \int_{a(x,n)}^{a(x,n)+l(x,n)} p^{n+c-1}(1-p)^{x+d-1} \right\} \mathrm{d}p \geq 1 - \alpha. \tag{13}$$

3.2 *Average length criterion*

From Equations (4) and (5), the inequality of average length criterion (ALC) can be written as:

$$\sum_x \ell' \beta(n+c, x+d) \leq \beta(c, d)\ell, \tag{14}$$

where the length ℓ', corresponding to the HPD interval, is found for each given x and n by solving

$$\frac{1}{\beta(n+c, x+d)} \left\{ \int_{a(x,n)}^{a(x,n)+\ell'(x,n)} p^{n+c-1}(1-p)^{x+d-1} \right\} \mathrm{d}p = 1 - \alpha. \tag{15}$$

3.3 *Worst-outcome criterion*

The posterior variance has been mentioned in many articles as given by Equation (12); it is a function of c, d, y and n.

So, from Equation (12), setting $n+c+d=\gamma$, differentiation with respect to x yields

$$2x^2 + (\gamma + 3d + 1)x + (3\gamma + 2)d - \gamma(\gamma + 1) = 0.$$

So, for any fixed c, d and ℓ, the worst-outcome criterion (WOC) is satisfied by choosing the minimum n such that

$$\inf_{x \in X} \left\{ \frac{1}{\beta(n+c, x+d)} \left\{ \int_{a(x,n)}^{a(x,n)+\ell(x,n)} p^{n+c-1}(1-p)^{x+d-1} \right\} \mathrm{d}p \right\} \geq 1 - \alpha, \tag{16}$$

and taking

$$x = \frac{\gamma(\gamma - 3d + 1) - 2d}{\gamma + 3d + 1} \tag{17}$$

$$n > (d - c) + (d - 1). \tag{18}$$

4. General strategy for finding sample size

Computer algorithms were devised to attain the sample sizes relatively quickly. Each algorithm is composed of several sub-algorithms. A description of the main sub-algorithms is given below, followed by an outline of the steps required for each criterion.

4.1 *Algorithms objectives*

All algorithms employ a bisectional search strategy to arrive at the final sample size, which stops when criterion is satisfied for n but not for $n-1$. The lower bracketing limit was always assumed to be 0, whereas the upper limit was taken from the standard frequentist formula for geometric sample sizes. For each possible value of n, the relevant criterion was evaluated, and the next candidate was chosen depending on the result of the previous.

4.2 *Finding lower and upper limits of HPD intervals*

Although the previous paragraph indicates how integrals with known limits can be evaluated, another frequent problem was to find the particular limits corresponding to HPD intervals. The method of solution depends on whether the beta density is unimodal or monotone increasing or decreasing. It also depends on whether a and l are both unknown or whether l is given. In the latter case, lower (a) and upper ($a+l$) limits for unimodal densities can be found by solving the equation

$$\psi(a \mid x, n, c, d) - \psi(a + l \mid x, n, c, d) = 0,$$

where ψ is given by Equation (10). The case of a and l both unknown is two-dimensional but can be approached through a combination of techniques already mentioned. Good starting points for the search are helpful in reducing the computing time. For example, first approximations for a and l can often be obtained from the symmetric credible set, for which a is simply the lower $\alpha/2$ percentile of the appropriate beta density, and $a+l$ is the upper $\alpha/2$ point.

4.3 *Algorithm for ACC*

For the ACC, c, d, α and l are fixed constants and the coverage depends on x. The left-hand side of inequality (13) must be calculated and compared with the desired average coverage $1-\alpha$. For each value of x, $x = 0, 1, 2, \ldots$, in the sum, the upper and lower limits, which depend only on a, as well as the resulting definite integral can be calculated as indicated earlier. The sum is compared with $1-\alpha$, and the process continues until convergence.

4.4 *Algorithm for ALC*

For the ALC, c, d, α and l are again fixed constants and the length of the HPD interval depends on x. The minimum n such that Equation (14) is satisfied is sought. The vector $f(x)$, $x = 0, 1, 2, \ldots$, is calculated by using Equation (9). The length of the HPD interval, represented by the vector $\ell'(x, n)$, $x = 0, 1, 2, \ldots$, is found by taking the difference between the upper and lower limits found as indicated above in the case that both a and l are unknown. The result can then be compared with $1-\alpha$, and the search continuous until convergence.

4.5 *Algorithm for WOC*

For the WOC, c, d, α and l are fixed, with no averaging required. The value of x^* is determined by condition (17), if we accept this conjecture. Thus, for each n, we need only to calculate the left-hand side of inequality (16), this integral being calculated as indicated earlier.

5. Numerical results

It is important to note that various criteria produce wildly divergent estimates either from uniform prior distribution($c = 1, d = 1$) or non-uniform prior distribution as well. The ACC and ALC are averages, and it might be of interest to observe the individual lengths or coverages from which the average is calculated, though using computer simulation for $n, 1 \le n < 500$, the ACC turned out to be a difficult method to result in the sample size required except for $1-\alpha = 0.05, n > (d-c)+(d-1)$, throughout computer simulation for the three sample size determination methods, the uniform prior distribution ($c = 1, d = 1$) was used. To illustrate how ACC method was not an acceptable method to search for sample size required, the authors considered the beta distributions β (3, 2), β (4, 3), β (5, 5), β (6, 4), β (7, 5), β (10, 5) and β (10, 7), and investigation resulted that no sampling is required for different coverages 0.95, 0.90 and 0.50. Different choices of a and l were considered, and Table 1 shows some specific values that lead to the sample sizes presented.

Table 1. Values of n_{ACC} for different β distributions.

$\beta(c, d)$	a	ℓ	n
$\beta(1, 1)$	0.4	0.15, 0.2, ..., 0.55	2
$\beta(3, 2)$	0.4	0.1, 0.15, ..., 0.55	2
$\beta(4, 3)$	0.4	0.45, 0.5, ..., 0.85	3
$\beta(5, 5)$	0.4	0.1, 0.15, ..., 0.55	6
$\beta(6, 4)$	0.4	0.15, 0.2, ..., 0.55	3
$\beta(7, 5)$	0.4	0.2, 0.25, ... 0.55	4
$\beta(10, 5)$	0.4	0.1, 0.15, ..., 0.55	1
$\beta(10, 7)$	0.1	0.55, 0.6, ..., 0.85	5

Table 2. Different values of n_{ALC} for $\beta(1,1)$.

$1-\alpha$	a	ℓ	n
0.50	0.2	0.01	2
		0.02	27
		0.03	42
		0.04	70
0.90	0.22	0.01	56
		0.04	3
		0.08	2
0.95	0.22	0.01	49
		0.05	3
0.99	0.22	0.01	44
		0.1	2
		0.22	1

Table 3. Different values of n_{WOC} for $\beta(1,1)$.

$1-\alpha$	a	ℓ	n	x
0.50	0.01	0.9	1	0
		0.8	1	0
		0.7	1	0
		0.6	2	1
		0.5	52	48
0.90	0.01	0.9	2	1
		0.8	5	3
		0.7	11	8
		0.6	28	25
0.95	0.01	0.9	4	2
		0.8	6	4
		0.7	14	11
		0.6	45	41
0.99	0.01	0.9	5	3
		0.8	11	8
		0.7	22	19
		0.6	75	71

The ALC criterion yields the sample size required under certain choices of a and l when l approaches zero (Table 2).

It is evident that WOC maintained the minimum sample size required and turned out to be the effective method for use in our approach of finding sample size provided a but not l is close to zero (Table 3).

6. Example and practical implementation

Pielou [11] studied two species of trees *Pinus ponderosa* I and *Pseudotsuga menziesu* II which together form an open woodland on the lower mountain slopes on either side of Okanogan valley in British Columbia. Species II grew in the woodland of species I. The observed distribution of run lengths of species II with respect to species II were obtained in Table 1 in Pielou's paper [11] to show that pairs of species exhibited segregation.

Table 4. Difference between n_{ALC} and n_{WOC}.

$1-\alpha$	a	ℓ	n_{ALC}	n_{WOC}	x
0.50	0.01	0.55	5	23	22
		0.59	3	10	10
		0.62	2	5	6
	0.1	0.33	9	79	76
		0.41	3	7	8
		0.44	2	5	4
0.90	0.01	0.64	3	23	22
0.95	0.2	0.24	4	40	38
	0.3	0.02	5	55	52
		0.04	4	39	37
0.99	0.01	0.65	3	45	43
	0.04	0.69	1	14	14

Table 5. Different values of n_{ALC} and n_{WOC} for Pielou's data.

$1-\alpha$	a	ℓ	n_{ALC}	n_{WOC}	x
0.50	0.01	0.53	15	115	105
		0.56	9	54	48
0.90	0.09	0.42	10	113	103
0.95	0.1	0.46	4	53	47
		0.5	2	29	26
	0.09	0.48	4	53	47
		0.52	2	29	26
0.99	0.2	0.25	5	88	79

In searching for biological model to account for this, Pielou [11] showed that our Bayesian model under consideration fitted the observed distributions of run lengths and estimated the parameters from the data giving $c = 10.0697$, $d = 3.2221$ (Table 4).

In this special case, ACC required no sampling through all values of $1-\alpha = 0.50$, 0.90, 0.95 and 0.99, whereas ALC yields a sample size estimate much less than the one obtained by WOC showing an extremely huge difference between both sample size estimates (Table 5).

Another example is the distribution of run lengths of the species I '*Geum triflorum*' growing in the transects of the species II '*Zygadenus venenosus*', as shown by Table 3 in Pielou's paper [11]. The parameter estimates of *G. triflorum* were calculated from the data to be $c = 23.8144$ and $d = 8.5773$.

Again, examining the required sample size using the predescribed methods, ACC showed that no sampling is required, whereas ALC indeed gave a sample size much less than that given by WOC.

It is noteworthy, however, that in our case for a geometric variate with a parameter given by beta distribution, ACC turned out to be a not-so-good method to obtain the sample size required when ALC shows a considerable difference better than WOC to estimate sample size.

7. Conclusion

This paper demonstrates that methods of sample size determination for geometric sampling vary extensively and the statistician therefore has to choose which method to employ when calculating

sample size taking into consideration not only the parameters of the beta prior, but also the lower and upper limits of the HPD interval. When working with ACC, then there may be no need to sample. In such situations, the budget would determine the sample size.

References

[1] C.J. Adcock, *A Bayesian approach to calculating sample sizes for multinomial sampling*, Statistician 36 (1987), pp. 155–159.

[2] C.J. Adcock, *A Bayesian approach to calculating sample sizes*, Statistician 37 (1988), pp. 433–439.

[3] C.J. Adcock, *Bayesian approaches to the determination of sample sizes for binomial and multinomial sampling-some comments on the paper by Pham-Gia and Turkkan*, Statistician 41 (1992), pp. 399–404.

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

[5] C.J. Adcock, *The Bayesian approach to determination of sample sizes-some comments on the paper by Joseph, Wolfson and Du Berger*, Statistician 44 (1995), pp. 155–161.

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

[7] L. Joseph, D.B. Wolfson, and Du Berger, *Sample size calculations for binomial proportion via highest posterior density intervals*, Statistician 44 (1995), pp. 143–154.

[8] H. Pezeshk, *Bayesian techniques for sample size determination in clinical trials*, Stat. Methods Med. Res. 12 (2003), pp. 489–504.

[9] T. Pham-Gia, *Sample size determinations in Bayesian statistics*, Statistician 44 (1995), pp. 163–166.

[10] T. Pham-Gia and N. Turkkan, *Sample size determinations in Bayesian analysis*, Statistician 41 (1992), pp. 389–392.

[11] E.C. Pielou, *Runs of one species with respect to another in transects through plant populations*, Biometrics 18 (1962), pp. 579–593.

[12] S.K. Sahu and T.M.F. Smith, *A Bayesian method of sample size determination with practical applications*, J. Roy. Stat. Soc. A 169 (2006), pp. 235–253.

[13] J. Stamey and R. Gerlach, *Bayesian sample size determination for case-control studies with misclassification*, Comput. Statist. Data Anal. 51 (2007), pp. 2982–2992.

[14] J. Stamey, D. Young, and T. Bratcher, *Bayesian sample-size determination for one and two poisson rate parameters with application to quality control*, J. Appl. Stat. 33 (2006), pp. 583–594.