Sample Size Requirements For Stratified Random Sampling Of Pond Water With Cost Considerations Using A Bayesian Methodology

A.A. Bartolucci[a], A.D. Bartolucci[b], and K.P. Singh[c]

[a]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama 35294-0022 USA, Email: Albartol@uab.edu, Tel: 205-934-4906, FAX: 205-975-2540

[b]Department of Psychology, University of Georgia, Athens, Georgia 30602 USA

[c]Department of Biostatistics, School of Public Health, University of Texas Health Science Center at Forth Worth, Forth Worth, Texas 76107-2699 USA

Abstract: Estimating average environmental pollution concentrations and it=s variance is a fairly straight forward task in stratified random sampling. A more challenging concept is the introduction of the cost factor into this environmental model. Traditional statistical techniques have incorporated costs from sampling within a stratum as well as stratum weights to determine the stratum size and overall required sample size. Information in the form of informative prior distributions to determine a more coherent variance in the system yield a more precise Bayesian approach to the sample size and cost calculations. This approach results in a more efficient sampling strategy in terms of cost when considering a pre specified margin of error for the sampling mean as well as the more complicated situation of correlation among the strata samples.

K*eywords:* *Stratified; random sampling; cost; Bayesian*

1. INTRODUCTION

The Traditional statistical approaches to calculating overall and stratum sample sizes in a stratified random sample are fairly straight forward. The procedure is somewhat complicated with the incorporation of cost as well as the possibility of correlation among the stratum samples. Applications of such approaches employing several monitoring strategies are well known (Thornton et. al, 1982, Nelson and Ward, 1981, Reckhow and Chapra, 1983, and Gilbert, 1987). Our focus here is to consider a pond water environment in which the strata are basically depth levels. Weighting of the strata as well as the overall variance of the sample mean are the main components in our derived statistics to determine sample size within the stratum. The three situations considered are that of pre specified margin of error, pre specified fixed cost and correlation among the strata samples. Cost efficiency is seen for most situations with the introduction of Bayesian methodology (Dayal and Dickey, 1976, Bartolucci and Dickey, 1977, Birch and Bartolucci, 1983 and Bartolucci et. al., 1998). The thrust of the Bayesian approach is through the derivation of the posterior estimate of the variance derived from coherent inference on a normal variance in the Behren=s Fisher context (Dayal and Dickey, 1976, Bartolucci et. al., 1998). Comparisons of the traditional or classical and Bayesian methodologies are presented using summary data from determining the phosphorous concentration in a pond water sampling environment.

2. TRADITIONAL SETUP

Let N=total number of population units in the target population. N_h is the number of population units within each of the h stratum, h=1,.....L.

Clearly $N=\sum_{h=1}^{L} N_h$. With reference to the sample, n=total number of sampling units in the target sample. Likewise as above, $n= n_1 + n_2 + ...+ n_L = \sum_{h=1}^{L} n_h$. We define the weight of the stratum, h,

as $W_h=N_h/N$. The mean, μ, of the population of N units is:

$$\mu = (1/N)\sum_{h=1}^{L} N_h \mu_h = \sum_{h=1}^{L} W_h \mu_h \quad (1)$$

where μ_h is the mean of the h stratum and is estimated by

$$m_h = (1/n_h)\sum_{i=1}^{n_h} x_{hi} \quad (2)$$

where x_{hi}= ith observation in stratum, h. An unbiased estimate of u is

$$m_{st} = \sum_{h=1}^{L} W_h m_h . \quad (3)$$

Let $N_h/N = n_h/n$ in all strata, then

$$m_{st} = \sum_{h=1}^{L} \frac{n_h}{n} m_h = (1/n)\sum_{h=1}^{L}\sum_{i=1}^{n_h} x_{hi} . \quad (4)$$

We define $Var(m_h) = (1/n)\sum_{h=1}^{L} W_h \sigma_h^2$ where σ^2 is the variance of the h stratum. We estimate the stratum variance by $s_h^2 = (\frac{1}{n_h - 1})\sum_{i=1}^{n_h} (x_{hi} - m_h)^2$. It can be shown that for large N,

$$s^2(m_{st}) = \sum_{h=1}^{L} \frac{W_h^2 s_h^2}{n_h} . \quad (5)$$

It will be important to note the robustness of this sample mean variance in the Bayesian context.

3. COMPUTING THE OPTIMUM n

We give a brief overview of three methods to compute the optimum n.

i) Pre Specified Margin of Error (PMOE)

Letting $d=|m_{st}-\mu|$, we denote d as the pre specified margin of error, (Gilbert, 1987). The value d is such that

$$P(|m_{st}-\mu| \exists d) = \alpha \quad (6)$$

for small α. The optimum n (Cochran, 1977) is thus

$$n = \frac{z_{1-\alpha/2}^2 \sum_{h=1}^{L} W_h s_h^2 / d^2}{1 + z_{1-\alpha/2}^2 \sum_{h=1}^{L} W_h s_h^2 / d^2 N}, \quad (7)$$

where for N64,

$$n = z_{1-\alpha/2}^2 \sum_{h=1}^{L} W_h s_h^2 / d^2 \quad (8)$$

and $z_{1-\alpha/2}$ is the usual100(1-α/2) critical value of the standard normal distribution. Thus the optimum n_h for the hth stratum is $n_h = n W_h S_h / \sum_{h=1}^{L} W_h S_h$.

ii) Pre Specified Fixed Cost

We define the overall cost of the sampling as

$$Cost = C = C_o + \sum_{h=1}^{L} C_h n_h \quad (9)$$

where c_h is the cost per population unit in the hth stratum and c_o is the fixed overhead cost. This is a standard cost representation. Thus the optimum n can be derived as (Aczel,1999),

$$n = \frac{(C - C_o)\sum_{h=1}^{L} W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}} \quad (10)$$

As above the optimum n_h per stratum is

$$n_h = n W_h S_h / \sum_{h=1}^{L} W_h S_h \quad (11)$$

One can examine equation (10) in terms of its sensitivity to changes in the PMOE. Let $W_h=n_h/n$.

Then (10) can be rewritten as

$$n = \frac{(C - C_o)\sum_{h=1}^{L} n_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} n_h S_h \sqrt{c_h}} \quad (12)$$

If we assume unequal PMOE, d_h , for sampling within stratum then we can write $n_h=(Z_{1-\alpha/2}S_h/d_h)$. (See Cochran, 1977, Aczel, 1999). Thus equation (12) can now be examined with respect to sensitivity to changes in d_h.

iii) Correlation among Depth Stratum

Let ρ_c = average correlation among all possible lags in the depth sampling environment. For example if L is the number of strata or depths and ρ_l=the correlation of the lth lag, then

$$\rho_c = (1/L)\sum_{l=1}^{L-1} \rho_l . \quad (13)$$

If n_h is the number to be sampled in each of the L strata or n_h = stratum size, then

$$n_h = [z^2_{1-\alpha/2} \sum_{h=1}^{L} W_{h s_h^2} / d^2][1 + \rho_c (L-1)] / L. \quad (14)$$

4. BAYESIAN CONSIDERATIONS

Examining equations (7), (10), (12) and (14) we see that they all involve the expression for the stratum variance, s^2_h. We reevaluated these expressions adding a prior structure to the variance (Dayal and Dickey, 1977, Bartolucci et. al. 1998) and then estimating the posterior expression for the variance, normal σ^2. We assumed an underlying normal distribution with both mean, μ, and variance, σ^2 unknown. In this context we define the likelihood function for n observations:

$$l(\mu,\sigma) \propto \sigma^{2(n/2)} \exp[-\frac{1}{2}\sigma^2 (n(\mu-m)^2 + \nu s^2] \quad (15)$$

for v=n-1, $nm=x_1+x_2+.....+x_n$, $vs^2=(x_1-m)^2 + (x_2-m)^2 +.....(x_n-m)^2$ and $\propto$ denotes a proportional relationship. Consider the t-density,

$$\phi(x; s^2) = s^{-1}[\upsilon^{1/2} Beta(\upsilon/2, 1/2)]^{-1} (1 + \upsilon^{-1}(x/s)^2)^{-(\upsilon+1)/2} \quad (16)$$

where $$Beta(a,b) = \int_0^1 z^{a-1}(1-z)^{b-1} dz$$

and υ, s>0.

The prior for μ is

$$p(\mu) = \phi_{\upsilon_o}(\mu - m_o, s_o^2) \quad (17)$$

for υ_o64.

The prior for σ^2 is

$$p(\sigma^2) \propto \tau g^2 / \chi^2_\tau, \quad \tau>0, g>0 \quad (18)$$

Where χ^2_τ is chi square on τ degrees of freedom. Thus considering expressions (15), (17), and (18) the posterior variance is

$$\varepsilon^2 = (\upsilon s^2 + \tau g^2)/B \quad (19)$$

where $B=\upsilon+\tau$.

Thus substituting ε_h^2 for s_h^2 in (7), (10), (12), and (14) yields the Bayesian estimates of n and n_h. Thus in the following section we apply the Bayesian analysis to these expressions to demonstrate and determine the efficiency of these expressions in terms of the sample size requirements and cost of sampling.

5. EXAMPLE

We wish to estimate the average phosphorous concentration (μg/100 ml) in pond water. The concentration of 100 ml aliquot from each 1 liter sample will be measured. The statistics for a classical representation of the data using the pre specified margin of error (PMOE) are given in Table 1. There are 5 depth strata to the pond in which N=total number of 100 ml water samples in the pond. N_h is the number of aliqots in stratum h. The weights, number samples from each strata, mean and variance of each strata are given as well all derived from our previous formulations above. We have assigned costs to each strata. For the sake of simplicity and without loss of generality we have reduced the costs to integer units. The cost for sampling stratum 1 and 2 are each 1. The costs assigned to strata 3, 4, and 5 are 2, 2, and 3 respectively - the assumption being that costs increase as the depth increases. Thus the overall cost of sampling is 74 units. Using the PMOE approach in Table 2 demonstrates the Bayesian results using empirical prior sampling information and incorporating that into the variance calculation overall. One sees that for realistic prior assignments of υ, τ and g in (13) that one realizes a reduction in assigned number per strata overall as well as a cost reduction. In Table 3 using pre specified overall cost did not yield any savings using the classical (top row) vs. the Bayesian approach (bottom row) this makes sense somewhat in that the cost is already fixed. However, we did examine these results using (12) in which we varied the PMOE , d_h , to determine the effect on cost using sensitivity changes and the classical and Bayesian results remained fairly equal. Table 4 summarizes the data results introducing correlation among the strata as per (14). The average correlation is in the first column. One can see that as you increase the average correlation , (13), then the required number sampled within each strata will increase, but at a slower rate in the Bayesian context.

Overall it appears that: Compared to the classical sampling analysis for the pre specified margin of error approach as well as the correlational approach, the Bayesian analysis resulted in a reduction in required samples thus lowering the cost, especially when realistic (empirical) prior hyperparameters are utilized. Also there was no serious impact on the posterior standard error of the estimates of the mean concentration. However, there were no real differences between the classical and Bayesian approaches in the pre specified fixed cost analysis. Given the current computational tools the Bayesian calculations proved to be fairly straight forward. Also given the current availability of databases, future Bayesian approaches to environmental sampling should be given serious consideration especially where costs are concerned.

REFERENCES

Aczel, A.D.,. Complete Business Statistics, McGraw Hill, Boston, (1999).

Bartolucci, A.A. and Dickey, J.M, Comparative Bayesian and traditional inferences for Gamma modeled survival data. *Biometrics*, **32**(2),.343-354, (1977).

Bartolucci, A.A., Blanchard, P.D., Howell, W.M., and Singh, K.P., A Bayesian Behren=s Fisher solution to a problem in taxonomy, *Environmental Modeling and Software*, **13,** 25-29, (1998).

Birch, R. and Bartolucci, A.A., Determination of the hyperparameters of a prior probability model in survival analysis, *Computer Programs in Biomedicine*, **17,** 89-84, (1983).

Cochran, W.G., *Sampling Techniques*,. Wiley Pub., 3rd edition, New York, (1977).

Gilbert, R.O., *Statistical Methods For Environmental Pollution Monitoring*, Van Nostrand Pub., New York, (1987).

Nelson, J.D. and Ward, R.C., Statistical considerations and sampling techniques for ground-water quality monitoring, *Ground Water*, **19,** 617-625, (1981).

Reckhow, K.H. and Chapra, S.C., Engineering Approaches for Lake Management, Volume 1, *Data Analysis and Empirical Modeling*, Butterworth, Boston, (1983).

Thornton, K.W., Kennedy, R.H., Magoun, A.D. and Saul, G.E., Reservoir water quality sampling design. *Water Resources Bulletin*, **18**, 471-480, (1982).

Table 1. Data for stratified random sampling to estimate samples per strata (PMOE)
Classical Approach (υ =1, τ =0, g =1) $s^2(m_{st})$ = 0.0140, Cost=74

Strata	N_h	W_h	n_h	m_h	s^2_h
1	4.25M	0.266	10	1.67	0.4376
2	3.96M	0.248	9	2.83	0.4228
3	3.23M	0.202	8	3.59	0.5339
4	2.85M	0.178	9	4.23	0.7222
5	1.70M	0.106	7	5.31	1.3920
Total	15.99M	1.000	43	-	-

Table 2. Bayesian Results (PMOE)

(υ,τ,g)	n_1	n_2	n_3	n_4	n_5	Total	$s^2(m_{st})$	Cost
35,1,0.5	9	9	8	8	7	41	0.0140	71
35,2,0.5	9	8	8	8	7	40	0.0141	70
20,1,1.0	9	8	8	8	6	39	0.0138	67
40,35,0.2	5	5	4	4	4	22	0.0143	38
40,35,0.5	7	7	6	6	4	30	0.0195	42

Table 3. Pre specified fixed cost

$C\text{-}c_0$	υ	τ	g	n	n_1	n_2	n_3	n_4	n_5
50	-	-	-	31	7	7	6	6	5+
50	40	35	0.12	31	7	7	6	6	5

Table 4. Example Using the Correlation Structure, ρ_c.

prior (υ,τ,g)	Classical		(35,1,0.5)		(20,1,0.1)		(40,35,.2)	
ρ_c	n_h	Cost	n_h	Cost	n_h	Cost	n_h	Cost
0.05	10	90	10	90	10	90	05	45
0.10	12	108	12	108	11	108	06	48
0.15	14	126	13	117	13	117	07	63
0.25	17	153	16	144	16	144	09	81
0.35	21	189	20	180	19	171	11	99
0.45	24	216	23	207	22	176	12	108
0.55	28	252	26	234	25	225	14	126