Sample Size Requirements for Stratified Random Sampling of Agricultural Run Off Pollutants in Pond Water with Cost Considerations Using a Bayesian Methodology

A.A. Bartolucci

Department of Biostatistics, University of Alabama at Birmingham, Birmingham, Alabama 35294-0022 USA

S.J. Bae and K.P. Singh

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 from fertilization components and their 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.

Keywords: *Stratified; random sampling; cost; Bayesian; margin of error; optimum*

1.0 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 as in 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 ations with the introduction of Bayesian methodology developed by Dayal and Dickey (1976), Bartolucci and Dickey (1977), Birch and Bartolucci (1983), Baldi and Long (2001) 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

of 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.

The motivation is to assure that we have a design that conforms to cost effectiveness guidelines recommended by the National Academy of Sciences (1977,2004) and Bartram and Balance (2001). These chosen designs incorporating a cost analysis will either achieve a specified level of effectiveness at minimal cost or a specified effectiveness at a specified cost. The incorporation of the Bayesian analysis as modeling the strata variance allows a further cost savings in the overall approach. The approach can be applied to sampling contaminants from well water or pond water with special attention to agricultural runoff as seen in Atzeni. Casey and Skerman (2001). Also Gilbert et. al. (1975) weighed in on the importance of this approach when cost considerations demanded attention when sampling radioactive pollutants from desert sites in Nevada. Our proposed technique can be applied to the sampling plans of Ward, R.C., Loftis, J.C. and McBride, B.G. (1990) as well as others. Thus historically there are many applications requiring the cost considerations as well as can be refined by cost considerations when sampling from the environment.

In section 2.0 below we derive the traditional set up of the sampling providing the basic statistics such as the sampling mean, variance, depth stratum size and weights as well as the overall population size. In section 3.0 we incorporate into our formulation the methodology for computing the optimum sample size under the assumptions of the pre specified margin of sampling error (PMOE). We then introduce cost consideration into the approach at a pre specified fixed cost per stratum for independent stratum as well as correlated stratum. In section 4.0 we introduce the Bayesian considerations in our methodology, especially as applied to the stratum variance which impacts on the overall final cost. In section 5.0 we apply the method to an example when sampling phosphorous concentration in pond water at 5 depth strata and demonstrate the conditions of cost reduction with the Bayesian methodology.

2.0 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.

3.0 COMPUTING THE OPTIMUM n

An important aspect of stratified random sampling is to determine how many samples are to be collected within a stratum. Gilbert (1987) has proposed a method for doing so that will minimize the variance $s^2(m_{st})$ in equation (5) above for a pre specified fixed cost per stratum or that will minimize the value of $s^2(m_{st})$ under the condition of a pre specified margin of error (PMOE). The PMOE is the value d such that $d=|m_{st}-\mu|$ or the minimal absolute distance we wish to tolerate between the sample mean and population mean with some acceptable error which we define below. We also relate these two conditions in our development of 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 as in Gilbert (1987). The value d is such that

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

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

$$n=\frac{z^2_{1-\alpha/2}\sum_{h=1}^{L}W_h s_h^2/d^2}{1+z^2_{1-\alpha/2}\sum_{h=1}^{L}W_h s_h^2/d^2 N}, \tag{7}$$

where for $N\rightarrow\infty$,

$$n=z^2_{1-\alpha/2}\sum_{h=1}^{L}W_h s_h^2/d^2 \tag{8}$$

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 $n_h=nW_h s_h/\sum_{h=1}^{L}W_h s_h$.

ii) Pre Specified Fixed Cost

We define the overall cost of the sampling as

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

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 in Aczel (1999),

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

As above the optimum n_h per stratum is

$$n_h=nW_h s_h/\sum_{h=1}^{L}W_h s_h \tag{11}$$

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

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

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) and 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

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

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

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

4.0 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 of 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:

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

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,

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

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

and υ, s>0.

The prior for μ is

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

for $\upsilon_o \to \infty$.

The prior for σ^2 is

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

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

$$\varepsilon^2 = (\upsilon s_h^2 + \tau g^2)/B \tag{19}$$

where B=υ+τ.

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.

5.0 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) d=0.2are given in Table 1. The PMOE=0.2 is a fairly reasonable choice in environmental sampling (see Gilbert, 1987). 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. Note that we have left N_h as a non integer just for the sake of generalization as this could be a depth measurement or volume or any convenient measure the sampler wishes to use or is convenient or has some environmental application. The weights are $W_h = N_h/N$ for each strata. The number samples from each strata, mean and variance of each strata are given as well all derived from our previous formulations above in section 3.0. 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. For example for the first strata we have n_1 =10 or 10x1=10 as a cost for the first strata and we have for the last strata n_5=7 at a cost of 3 per sample in that stratum or 7x3=21 for the cost of sampling that stratum. Thus doing likewise for the rest of the strata, we have a total cost of 74. Using the PMOE approach in Table 2, setting d=0.2 demonstrates the Bayesian results using empirical prior sampling information and incorporating that into the variance calculation overall. See equation (19). One sees that for realistic prior assignments of υ, τ and g in (19) and incorporating that variance into the calculation for n_h in section 3.0 one realizes a reduction in assigned number per strata overall as well as a cost reduction in Table 2. In Table 3 using pre specified overall cost (i.e. holding C constant in (9)) 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 (results not shown here). 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.

6.0 Discussion

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.

The importance of incorporating the correct elements into an environmental study design has been emphasixed by many authors. For example Smith (1984) discusses the efficiency of the design. In our case efficiency not only means precision in terms of the pre specified margin of error, but also on the cost considerations. Provost (1984) has also touched upon several of the elements discussed in this paper. These papers plus others examine the consequences of parameter estimation in terms of efficiency as one varies both the type of design and the size of the sampling effort. The stratified random sampling scheme discussed above is a useful and flexible design for estimating environmental concentrations, inventories and cost. They make use of prior information in the classical statistical sense of dividing the population into subgroups or

strata that are basically internally homogeneous. We have extended that prior knowledge to the Bayesian application of making use of the distribution of the prior variation within the strata to establish a more efficient design in terms of number sampled within the strata as well as cost efficiency. See Reckhow and Chapra (1983) for a further discussion of empirical modeling and data analysis in a stratified setting. Thus, have extended the work of several authors by using the Bayesian methodology to ensure not only an efficient design in the sampling sense, but in the cost arena as well. A possible opportunity for extension of this methodology is to consider multi stage sampling designs and the consequences of incorporating prior information into the variation components of primary as well as secondary units in the two stage setting. For designs with more complicated staging an extended multivariate model of the variation within the population can be considered.

7.0 REFERENCES

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

Atzeni, M., Casey, K., Skerman, A. (2001). A model to predict cattle feedlot runoff for effluent reuse applications. International Modeling and Simulation Society. Vol.4. pp1871-1876.

Baldi,P. and Long, A.D. (2001) A Bayesian framework for the analysis of microarray data: regularized t-test and statistical inferernces of gene changes. Bioinformatics., 17(6) , 509-519.

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

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

Bartram, J and Balance, J. (2001). Water Quality Monitoring: A Practical Guide to the Design and Implementation of Fresh Water Quality Studies and Monitoring Programs. Spon Press. London.

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

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

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

Gilbert, R.O. , Eberhardt, L.L., Fowler, E.B., Romney, E.M., Essington, E.H. and Kinnear, J.E. (1975). Statistical analysis of Pu and Am contamination of soil and vegetation on NAEG study sites. The Radioecology of plutonium and other transuranics in desert environments, M.G. White and P.B. Dunaway. Eds. US Energy Research and Development Administration. NVO-153. Las Vegas, pp 339-448.

National Academy of Sciences (1977). Environmental Monitoring: Analytical Studies for the U.S. Environmental Protection Agency. Vol. IV. National Academy of Sciences, Washington, D.C..

National Academy of Sciences (2004). Analytical Methods and Approaches for Water Resources Project. Planning Panel on Methods and Techniques of Project Analysis, Committee to Assess the U.S. Army Corps of Engineers Methods of Analysis and Peer Review for Water Resources Project Planning, National Research Council. Washington, D.C.

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

Provost, L.P. (1984). Statistical methods in environmental sampling. In Environmental Sampling for Hazardous Wastes. G.E. Schweitzer and J.A. Santolucito, eds. ACS Symposium Series 267. American Chemical Society , Washington, DC. pp 79-96.

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

Smith, W. (1984). Design of efficient environmental surveys over time. In Statistics in the Environmental Sciences, American Society for Testing and Materials, STP 845. S.M. Gertz and M.D. London, eds. American Society for Testing and Materials , Philadelphia. pp 90-97.

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

Ward, R.C., Loftis, J.C., and McBride, G.B. (1990). Design of water quality Monitoring systems. John Wiley and Sons. New York. Boston.

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	4.25	0.266	10	1.67	0.4376
2	3.96	0.248	9	2.83	0.4228
3	3.23	0.202	8	3.59	0.5339
4	2.85	0.178	9	4.23	0.7222
5	1.70	0.106	7	5.31	1.3920
Total	15.99	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 (Bayesian results in bottom row)

$C\text{-}c_0$	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